

Single- Cell Omics: experimental workflow, data analyses and applications

Fengying Sun^{1†}, Haoyan Li^{2†}, Dongqing Sun^{3,4†}, Shaliu Fu^{3,5,6,7†}, Lei Gu^{8†}, Xin Shao^{2,9†}, Qinjin Wang^{8†}, Xin Dong^{3,4†}, Bin Duan^{3,5,6,7†}, Feiyang Xing^{3,4†}, Jun Wu^{10†}, Minmin Xiao^{1*}, Fangqing Zhao^{11*}, Jindong Han^{12*}, Qi Liu^{3*}, Xiaohui Fan^{2,9,13*}, Chen Li^{8*}, Chenfei Wang^{3,4*}, Tieliu Shi^{1,10, 14*}

¹ Department of Clinical Laboratory, the Affiliated Wuhu Hospital of East China Normal University (The Second People's Hospital of Wuhu City), Wuhu 241000, China;

² Pharmaceutical Informatics Institute, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China;

³ Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration (Tongji University), Ministry of Education, Orthopaedic Department, Tongji Hospital, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai 200082, China;

⁴ Frontier Science Center for Stem Cells, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China;

⁵ Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai 200082, China;

⁶ Research Institute of Intelligent Computing, Zhejiang Lab, Hangzhou 311121, China;

⁷ Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai 201210, China;

⁸ Center for Single-cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200001, China;

⁹ National Key Laboratory of Chinese Medicine Modernization, Innovation Center of Yangtze River Delta, Zhejiang University, Jiaxing 314103, China;

¹⁰ Center for Bioinformatics and Computational Biology, Shanghai Key Laboratory of Regulatory Biology, the Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai 200241, China

¹¹ Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China;

¹² Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Center for Quantitative Biology (CQB), Peking University, Beijing 100871, China;

¹³ Zhejiang Key Laboratory of Precision Diagnosis and Therapy for Major Gynecological Diseases, Women's Hospital, Zhejiang University School of Medicine, Hangzhou 310006, China;

¹⁴ Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, School of Statistics, East China Normal University, Shanghai 200062, China;

Received 07-Dec-2023; accepted 18-Apr-2024

†Contributed equally to this work

*Corresponding authors (Minmin Xiao, email: 1084173819@qq.com; Fangqing Zhao, email: zhfq@biols.ac.cn; Jindong Han, email: jackie.han@pku.edu.cn; Qi Liu, email: qiliu@tongji.edu.cn; Xiaohui Fan, email: fanxh@zju.edu.cn; Chen Li, email: cli@shsmu.edu.cn; Chenfei Wang, email: 08chenfeiwang@tongji.edu.cn; Tieliu Shi, email: tlshi@bio.ecnu.edu.cn)

1 **Abstract:**

2 Cells are the fundamental units of biological systems and exhibit unique development
3 trajectories and molecular features. Our exploration of how the genomes orchestrate
4 the formation and maintenance of each cell, and control the cellular phenotypes of
5 various organisms, is both captivating and intricate. Since the inception of the first
6 single-cell RNA technology, technologies related to single-cell sequencing have
7 experienced rapid advancements in recent years. These technologies have expanded
8 horizontally to include single-cell genome, epigenome, proteome, and metabolome,
9 while vertically they have progressed to integrate multiple omics data and incorporate
10 additional information such as spatial scRNA-Seq and CRISPR screening. Single-cell
11 omics represent a groundbreaking advancement in the biomedical field, offering
12 profound insights into the understanding of complex diseases, including cancers. Here,
13 we comprehensively summarize recent advances in single-cell omics technologies,
14 with a specific focus on the methodology section. This overview aims to guide
15 researchers in selecting appropriate methods for single-cell sequencing and related
16 data analysis.

17

18 **Keywords:** single cell, sequencing; genome, epigenome, proteomics, metabolomics,
19 multimodal, spatial transcriptomics, CRISPR screening

20

21 **CONTENTS**

22 **Introduction.....**.....**1**

23 **Chapter 1 Single-cell transcriptome sequencing.....****1**

24 *Overview of scRNA-seq.....*.....**2**

25 *Currently available scRNA-seq technologies*.....**5**

26 *Low-throughput scRNA-seq technologies.....*.....**5**

27 *High-throughput scRNA-seq technologies.....*.....**10**

28 *Strategy for developing high-throughput scRNA-seq*.....**10**

29 *Plate-based high-throughput scRNA-seq methods*.....**10**

30 *Microfluidics-based high-throughput scRNA-seq methods*.....**13**

31 *Combinatorial indexing-based high-throughput scRNA-seq technologies....*.....**17**

32 *Single nuclei RNA-sequencing (snRNA-seq)*.....**18**

33 *Computational methods for scRNA-seq data*.....**19**

34 *Data pre-processing.....*.....**19**

35 *Data processing*.....**22**

36 *Extended downstream analysis*.....**33**

37 *Applications of scRNA-seq.....*.....**37**

38 *Summary*.....**46**

39 **Chapter 2 Single-cell whole-genome sequencing****46**

40 *ScWGA methods*.....**47**

41 *High-throughput scWGS methods*.....**52**

42 *Microfluidic-based high-throughput scWGS methods*.....**52**

43 *Nanowell-based high-throughput scWGS methods*.....**53**

44 *Combinatorial indexing-based high-throughput scWGS methods*.....**54**

45	<i>Applications of scWGS in biomedicine</i>	55
46	<i>Summary</i>	58
47	Chapter 3 Single-cell epigenome sequencing	58
48	<i>Techniques for sequencing the single-cell epigenome</i>	58
49	<i>Methylation</i>	58
50	<i>Chromatin accessibility and nucleosome positioning</i>	61
51	<i>Histone modification and transcription factor binding</i>	62
52	<i>3D genome structure</i>	63
53	<i>Computational methods for single-cell epigenome data</i>	63
54	<i>Reads preprocessing, quality control, and quantification</i>	63
55	<i>Imputation</i>	64
56	<i>Clustering</i>	65
57	<i>Cell type annotation and trajectory inference</i>	67
58	<i>Differential analysis and features selection</i>	67
59	<i>Gene regulation inference</i>	68
60	<i>Multi-function analysis pipelines</i>	69
61	<i>Applications of single-cell epigenomes</i>	70
62	<i>Summary</i>	72
63	Chapter 4 Single-cell proteomics technology based on mass spectrometry	72
64	<i>MS-based SCP Workflow</i>	73
65	<i>Single-cell Isolation</i>	73
66	<i>Sample preparation</i>	73
67	<i>MS analysis</i>	74
68	<i>State-of-the-art SCP tools</i>	77
69	<i>Integrated tools</i>	77
70	<i>Easy-to-use Tools</i>	79
71	<i>Applications</i>	80
72	<i>Discussion and prospects</i>	82
73	<i>Summary</i>	84
74	Chapter 5 Single-cell metabolomics technology	84
75	<i>Research techniques in single-cell metabolomics</i>	85
76	<i>State-of-the-art technologies and methods in single-cell metabolomics field based on mass spectrometry</i>	85
77	<i>Data analysis in single-cell metabolomics based on Mass spectrometry</i>	92
78	<i>Applications</i>	92
79	<i>Summary</i>	93
81	Chapter 6 Single-cell multimodal sequencing technology	93
82	<i>Single-cell multimodal sequencing technology</i>	94
83	<i>Transcriptome + gDNA</i>	94
84	<i>Transcriptome + epigenome</i>	94
85	<i>Transcriptome + proteome</i>	96
86	<i>Techniques capturing more than two modalities</i>	97
87	<i>Multi-omics integration analysis</i>	98
88	<i>Categories of multimodal integration tools</i>	98

89	<i>Benchmarks for single-cell multimodal integration tools</i>	100
90	<i>Summary</i>	101
91	Chapter 7 Single-cell spatial transcriptomics technology101	
92	<i>Techniques for spatially resolved transcriptomics</i>	101
93	<i>Microdissection-based ST techniques</i>	102
94	<i>Barcoding-based ST techniques</i>	103
95	<i>Imaging-based ST techniques</i>	103
96	<i>Techniques for spatial multi-omics</i>	105
97	<i>Computational methods for spatial transcriptomics</i>	106
98	<i>Denoising to enhance the signal in spatial transcriptomics</i>	106
99	<i>Subcellular analysis for imaging-based ST data</i>	107
100	<i>Deciphering spatial distribution of cell types by integrating scRNA-seq</i>	109
101	<i>Spatial domain identification</i>	112
102	<i>Detection of spatially variable genes and gene expression patterns</i>	115
103	<i>Pseudo-time trajectory analysis</i>	119
104	<i>Cell-cell communication and gene-gene interaction</i>	120
105	<i>Integrative analysis of spatial data</i>	121
106	<i>Applications</i>	124
107	<i>Summary</i>	128
108	Chapter 8 Single-cell CRISPR screening technology129	
109	<i>Transcriptome-based scCRISPR-seq</i>	129
110	<i>Epigenome-based scCRISPR-seq</i>	130
111	<i>Multimodal scCRISPR-seq</i>	131
112	<i>The tools to analyze scCRISPR-seq data</i>	132
113	<i>Applications of scCRISPR-seq</i>	133
114	<i>Summary</i>	134
115	Epilogue	135

116 **Introduction**

117 The exploration of individual cells enhances our understanding of cellular diversity,
118 disease processes, and the organization of multicellular organisms. Technologies for
119 measuring biological systems at the single-cell level have made exciting advances and
120 are now at the forefront of research. Single-cell RNA sequencing (scRNA-seq)
121 technique allows the dissection of gene expression at single-cell resolution,
122 revolutionizing transcriptomic studies. Since its initial discovery in 2009, more than
123 60 scRNA-seq protocols have been developed so far (Table S1). The maturation of
124 scRNA-seq provides researchers with unique opportunities to catalog human cell
125 types, understand their development, variation between individuals, and unravel their
126 involvement in disease. With the rapid development of single-cell sequencing
127 technology and reduction of cost, this has been widely used to solve critical
128 biomedical problems.

129 The rapid development of scRNA-seq technology has facilitated the exploration
130 of other omics, including genomics, epigenome, proteomics, and metabolomics.
131 Novel technologies, such as multi-omics technology, spatial scRNA-seq, and CRISPR
132 screening, have also emerged to gain a comprehensive understanding of complex
133 cellular behavior through multi-omics data integration and the incorporation of
134 additional information. Figure 1 illustrates the expanded landscape of single-cell
135 sequencing technologies.

136 This paper will review the latest developments of single-cell omics technologies
137 from the following eight aspects: (1) Single-cell transcriptome sequencing; (2)
138 Single-cell whole-genome sequencing; (3) Single-cell epigenome sequencing; (4)
139 Single-cell proteomics technology; (5) Single-cell metabolomics technology; (6)
140 Single-cell multimodal sequencing technology; (7) Single-cell spatial transcriptomics
141 technology; (8) Single-cell CRISPR screening technology. The aims are to
142 systematically summarize and discuss in detail currently available single-cell omics
143 technologies, the computational approaches to decipher the single-cell dataset, and
144 their advantages, disadvantages, and applications.

145

146 **Chapter 1 Single-cell transcriptome sequencing**

147

148 The advent of single-cell RNA sequencing (scRNA-seq) technology is a
149 state-of-the-art technique for analyzing cellular complexity and heterogeneity,
150 providing a wealth of information across diverse scientific domains. The high
151 resolution of this technology makes it possible to discuss novel biological issues by
152 offering a unique opportunity to explore the transcriptional landscape of single cells.
153 This cutting-edge method, since the first scRNA-seq protocol was introduced in 2009,
154 has seen substantial advancements in method development (Figure 2). These
155 approaches incorporate essential improvements and modifications in single-cell
156 isolation, capture, reverse transcription, cDNA amplification, library preparation,
157 sequencing, and data analysis to enhance throughput and automation while decreasing
158 time and costs. This chapter presents a comprehensive overview of single-cell
159 transcriptome sequencing technologies, bioinformatics analysis methods, and their

160 uses in medical and biological sciences research to help researchers make informed
161 choices.

162

163 *Overview of scRNA-seq*

164 ScRNA-seq, the pioneered single-cell sequencing technology, has witnessed wide
165 popularity and encompasses a variety of approaches. Despite the diversity in methods,
166 they all follow a similar general process involving four primary steps: (i) isolation of
167 single cells, (ii) reverse transcription (RT), (iii) cDNA amplification, and (iv)
168 sequencing library construction and sequencing (Hedlund and Deng, 2018). Major
169 steps of the scRNA-seq workflow are shown in Figure 3. This section outlines some
170 techniques and solutions related to single-cell isolation and sequencing library
171 construction.

172

173 *Single-cell isolation*

174 In scRNA-seq, the first and critical step is **single-cell isolation** in which tissue
175 dissociation and single-cell separation are considered significant contributors to
176 contamination, batch effects, and procedural disparities (Tung et al., 2017). Thus, to
177 perform high-throughput and unbiased single-cell sequencing, a reliable and accurate
178 capturing of single cells with high efficiency is the key determinant. Early methods
179 for single-cell isolation including limited serial dilution (Gross et al., 2015), manual
180 micromanipulation (Hu et al., 2016a), and laser capture microdissection (LCM)
181 (Emmert-Buck et al., 1996) were low-throughput, time-consuming, inefficient, and
182 technically challenging but are still used to analyze low number of cells (e.g. rare cells)
183 (Dal Molin and Di Camillo, 2019).

184 Fluorescence-activated cell sorting (FACS), a commonly used high-throughput
185 technique, offers specific and automated isolation of thousands of individual cells but
186 requires a large input volume (numbers of cells for isolation >10,000) (Hu et al.,
187 2016a). Moreover, this technique is inadequate for certain cells exhibiting low marker
188 expression due to the faint or weak fluorescence signal making it challenging to
189 differentiate subpopulations with similar marker expression (Yasen et al., 2020).
190 Magnetic-activated cell sorting (MACS) is another high-throughput isolation
191 technique designed to separate various cell types based on enzymes, lectins,
192 antibodies, or streptavidin conjugated to magnetic beads, facilitating the binding of
193 specific proteins on the target cells (Hu et al., 2016a). The MACS system boasts a
194 notable advantage, achieving >90% purity for specific cell populations (Miltenyi et al.,
195 1990). However, MACS has inherent limitations compared to FACS due to
196 immunomagnetic techniques that can only isolate cells into negative and positive
197 populations. Moreover, it cannot isolate cells based on low or high expression of a
198 molecule, a capability present in FACS (Hu et al., 2016a). In the current landscape of
199 high-throughput sequencing platforms, methods involving microfluidic-based
200 single-cell manipulation have emerged as the leading technique for single-cell
201 separation in transcriptome studies and have significantly enhanced the scale,
202 efficiency, and accuracy of the isolation process. Microfluidics devices, in which
203 reaction chambers or droplets are used to capture the cells followed by individual

204 steps nanoliter reactions, offer a cost-effective and sample-efficient analysis. These
205 devices are primarily categorized into microwell-based methods, droplet-based
206 methods, and integrated fluidic circuits (IFCs). The integration of microfluidics
207 systems in scRNA-seq has significantly enhanced sequencing throughput, enabling
208 the simultaneous processing and analysis of tens of thousands of single cells. A
209 comprehensive overview of current single-cell isolation technologies, including their
210 advantages and limitations, is presented in Table S2.

211

212 *Reverse transcription and cDNA amplification*

213 A single mammalian cell contains approximately 10 picograms aggregate quantity of
214 RNA, with a predominant portion composed of ribosomal RNAs (rRNAs) and
215 transfer RNAs (tRNAs), whereas messenger RNAs (mRNAs) constitute only 1%~5%
216 of the total (Liu et al., 2014; Wang et al., 2023c). Since mRNA is present in extremely
217 low amounts in a single cell, it is essential to amplify cDNA after the reverse
218 transcription (RT) process to obtain significant quantities for sequencing library
219 preparation. cDNA amplification can be achieved through either exponential
220 amplification using polymerase chain reaction (PCR) or linear amplification based on
221 in vitro transcription (IVT).

222 Presently, the predominant method for library construction involves PCR-based
223 cDNA amplification, including poly(A) tailing and template switching (TS) methods
224 (Kolodziejczyk et al., 2015). The poly(A) tailing method employs an oligo-dT primer
225 that binds the mRNA 3'-poly(A) tail to reverse transcribe mRNA into cDNA. The
226 poly(A) tailing method is speedy but cannot capture nonpolyadenylated (Poly(A)-)
227 RNA (Hebenstreit, 2012), and the capture efficiency is low, which is reported to be
228 around 10%~15% for current protocols (Islam et al., 2014). Additionally, the
229 termination of the reverse transcriptase reaction may lead to diminished coverage rate
230 of the 5' end of mRNA in transcription. The TS technique involves the utilization of
231 Moloney murine leukemia virus (MMLV) reverse transcriptase for the TS process.
232 MMLV reverse transcriptase can append a poly(C) tail to the trailing end of the newly
233 synthesized single strand. The poly(C) tail can bind to the 5'-end poly(G) tail of
234 template switching oligonucleotide (TSO) adapter sequences. Following this
235 interaction, a “switch” takes place: the reverse transcriptase utilizes the TSO as a
236 template to synthesize cDNA, to complete the adaptor conversion (Picelli, 2017). The
237 TS method exhibits a reduced susceptibility to nucleic acid loss, but it comes with a
238 lower sensitivity compared to the poly(A) tailing method. PCR-based amplification,
239 while capable of rapidly and efficiently amplifying substantial quantities of cDNAs in
240 a short timeframe, is constrained by the inherent characteristics of the exponential
241 amplification process. PCR tends to favor the amplification of shorter and less
242 G-C-rich amplicons, resulting in quantification bias and the accumulation of
243 non-specific transcripts, leading to the loss of original transcript information (Aird et
244 al., 2011). Additionally, PCR can cause the over-predominance of highly expressed
245 transcripts in the final library (Aird et al., 2011).

246 IVT represents another method for cDNA synthesis and amplification. The key
247 strength of IVT is its linear amplification, a feature that mitigates what is commonly

known as amplification bias, rendering it considered more precise and reproducible in comparison to PCR (Grün and van Oudenaarden, 2015; Chen et al., 2017a). In this strategy, an oligo-dT primer, encompassing 1) a unique molecular identifier (UMI, featuring random nucleotide sequences that label individual mRNA molecules and are employed for quantifying unique transcripts and correcting amplification biases); 2) a unique cell barcode; 3) an Illumina adapter; and 4) a T7 promoter, initially binds to the 3'-poly(A) tail of mRNA to synthesize both strands of cDNA. Subsequently, the cDNA fragments, each uniquely barcoded, are combined, and the T7 polymerase identifies the T7 promoter sequence, initiating IVT to generate additional RNA molecules. Another round of RT is required to reconvert the amplified RNA molecules into cDNAs for sequencing library construction. Consequently, the resultant cDNAs exhibit a pronounced bias in coverage toward the 3' end. Moreover, this approach is characterized by its time-intensive and labor-intensive nature, contributing to its less widespread use compared to PCR (Hashimshony et al., 2012).

As a unique step of scRNA-seq, all transcripts originating from a single cell obtain a unique barcode-a brief oligonucleotide sequence introduced into the cDNA during RT to distinguish the transcriptome of individual cells. Leveraging the unique barcode information, these transcripts can be readily attributed to their respective cells. Consequently, a multitude of single-cell transcriptomes, each uniquely labeled with cell barcodes, are amalgamated for the library construction and subsequent sequencing in a single run. This approach significantly diminishes costs and augments sequencing throughput. The cell barcode serves as a highly effective strategy for enabling parallel processing and was initially used in the STRT-seq method in 2011 (Islam et al., 2011). Subsequently, in 2014, Islam and colleagues further innovated by introducing a UMI to identify each cDNA molecule within a cell (Islam et al., 2014). Like cell barcodes, a UMI is likewise a 6-8 base pair random oligonucleotide sequence that can be integrated into each transcript within a cell during RT. With a substantial pool of UMIs, each transcript is tagged with a distinct barcode, thereby ensuring that all duplicated molecules produced through PCR amplification retain the original transcript's UMIs. By counting the unique barcodes, the number of original copies of the transcript can be accurately quantified, thereby eliminating the bias of PCR amplification.

Regarding the three pivotal issues in single-cell RNA sequencing: (1) isolating individual cells; (2) minimizing RNA loss during RT; and (3) producing enough DNA for sequencing, researchers have undertaken explorations into technological advancements in recent years. This has led to the proposal of various scRNA-seq methodologies tailored for the comprehensive study of single-cell transcriptomics.

The different scRNA-seq technologies are rooted in shared fundamental principles, yet they vary in at least one of the following aspects: (1) single-cell isolation; (2) RT; (3) amplification of cDNA; (4) transcript coverage; (5) UMI; (6) strand-specificity. Each approach has its unique strengths and limitations. The key features of these diverse scRNA-seq technologies are consolidated in Table S1. Consequently, researchers have the flexibility to choose a suitable scRNA-seq method based on technical characteristics, advantages, cost factors, and throughput demands.

292

293 *Currently available scRNA-seq technologies*294 *Low-throughput scRNA-seq technologies*

295 Over 60 distinct approaches have been developed for scRNA-seq, as detailed in Table
296 S1. In general, these approaches can be categorized into two main types:
297 low-throughput and high-throughput methods. The fundamental chemistry developed
298 in low-throughput approaches is geared towards enhancing sequencing sensitivity and
299 accuracy, while reducing costs and technical noise. Particularly, sensitivity stands out
300 as the foremost critical feature, serving as a fundamental indicator of method
301 performance. High-throughput methods have evolved from the essential chemistries
302 of several classic low-throughput approaches. We will discuss these technologies in
303 detail below.

304

305 *Tang's protocol*

306 Tang's protocol, developed by the Surani group in 2009, stands as the pioneering
307 scRNA-seq method. In this technique, a microscope is used to manually select a
308 single mouse blastomere. Following that, the cells are lysed, and the mRNAs are
309 converted into cDNAs through RT using a poly(T) primer featuring an anchor
310 sequence (UP1). A poly(A) tail is then added to the trailing end of the first strand end
311 with the help of terminal transferase. Following this, the second strand cDNAs are
312 synthesized using the second poly(T) primer featuring another anchor sequence (UP2).
313 The cDNAs are then efficiently amplified through PCR using UP1 and UP2 primers,
314 and libraries are constructed for sequencing on the SOLiD system. This method can
315 generate nearly full-length cDNA of transcripts and detect ~13,000 genes (Tang et al.,
316 2009). This technology mainly helps in the unveiling of new transcripts and
317 alternative splicing isoforms.

318 Despite representing a significant advancement for the emerging field of
319 scRNA-seq at the time, this approach has considerable limitations. First, the approach
320 can only identify mRNA that has a poly(A) tail; it cannot capture mRNAs without
321 poly(A) tails, such as histone mRNAs, miRNA, circular RNA (circRNA), and nascent
322 RNA. This is because the method depends on poly(T) primers for the capturing of
323 mRNA via the poly(A) tail for initiating RT reactions. Second, inefficiencies in the
324 enzymatic reactions contribute to a reduction in sequencing sensitivity, resulting in
325 the loss of low-expression transcripts. Third, this method is not strand-specific and
326 cannot distinguish between sense and antisense transcripts. Thus, it is not commonly
327 used.

328

329 *STRT-seq*

330 In 2011, Islam et al. developed a single-cell tagged reverse transcription sequencing
331 (STRT-seq) method, a highly multiplexed approach for scRNA-seq on the Illumina
332 platform (Islam et al., 2011). This method introduced a barcode and an upstream
333 primer-binding sequence through the template-switching mechanism during RT,
334 facilitating strand-specific amplification of 3' ends and high throughput 96-cell
335 multiplexing. Using an array-based strategy, STRT-seq supports the processing of up

336 to 800 individual cells. A key advantage of this method is that it can add a unique
337 barcode sequence to each cell during the RT process, thereby enabling large-scale
338 detection of various mixed cell samples, such as highly heterogeneous tumor cell
339 samples. Compared to Tang's method, STRT-seq significantly reduces costs and
340 processing time through its early barcoding strategy. However, it employs multiple
341 cycles of PCR, potentially introducing PCR bias. This method has many applications
342 in the biomedical field, including the characterization of tumor heterogeneity and the
343 identification of potential novel biomarkers or drug targets for disease diagnosis and
344 treatment (Cui et al., 2021; Song et al., 2022a; Tian et al., 2022).

345

346 *Smart-seq*

347 In 2012, Ramsköld *et al.* developed Smart-seq, a reliable and consistent approach for
348 scRNA-seq, a switching mechanism at the leading end of the transcript, which
349 demonstrated significant improvement, surpassing 40% efficiency in full-length
350 cDNA synthesis for transcripts (Ramsköld *et al.*, 2012). The publication of the
351 technology is a landmark in the field of scRNA-seq studies. The core principle of
352 Smart-seq involves utilizing poly(T) primers and SMART-TS technology to convert
353 polyadenylated (poly(A)+) RNA into full-length cDNA. The resultant cDNA
354 molecules, after the amplification using PCR, are then used to create Illumina
355 sequencing libraries by the Nextera Tn5 transposome technique. This method
356 considerably improves the ability to detect alternatively spliced exons and
357 low-abundance expressed transcripts. Smart-seq method has been widely used in
358 medicine, such as analyzing gene expression profiles of CD177⁺ cells in the liver of a
359 mouse model with biliary atresia (Zhang et al., 2022d); to analyze the genome-wide
360 expression of skeletal muscle stem, niche cells, and single myofibers (Blackburn et al.,
361 2019; Blackburn et al., 2021); and investigating differences in dermal CD4⁺ Trm cells
362 between patients with acute cutaneous lupus erythematosus and normal controls
363 (Zhao et al., 2022c).

364 Smart-seq2 was developed to overcome the limitations of coverage, less
365 productivity, and sensitivity issues observed in Smart-seq (Picelli *et al.*, 2013). In
366 order to increase cDNA library yield and length, several improvements were
367 implemented in Smart-seq2 including enhancement of reverse RT, TSOs, and the
368 reamplification of PCR. A notable enhancement in cDNA yield, approximately
369 twofold, was achieved by incorporating a locked nucleic acid (LNA) guanylate at the
370 3' end of TSO as opposed to Smart-seq. This improvement can be attributed to the
371 heightened thermal stability of LNA-DNA base pairs. Moreover, the addition of the
372 methyl group donor betaine, along with a higher concentration of MgCl₂, also led to a
373 substantial increase in cDNA yield. Commencing the addition of deoxyribonucleoside
374 triphosphates (dNTPs) before RNA denaturation, as opposed to incorporating them in
375 the RT master mix, enhanced the average length of the preamplified cDNA. This
376 improvement is likely attributed to increased stability in the hybridization of RNA to
377 the oligo-dT primer. The use of KAPA HiFi Hot Start DNA polymerase improved
378 cDNA generation and achieved greater cDNA length. Smart-seq2 transcriptome
379 libraries outperform Smart-seq in terms of detection strength, coverage rate, bias, and

accuracy. Smart-seq2 transcriptome libraries can be generated with off-the-shelf reagents even at a lower cost, allowing the in-depth analysis of entire exons of each transcript and also detecting different splice variants. This method also facilitates thorough analysis of single nucleotide polymorphisms (SNPs) and mutations. However, it has limitations, such as the lack of strand specificity and the incapability to detect poly(A)- RNAs (Picelli et al., 2014). Additionally, the cell isolation process using micropipettes is time-consuming and low-throughput.

Smart-seq3 represents an enhancement in sensitivity achieved through the optimization of RT and TS conditions (Hagemann-Jensen et al., 2020). The optimal parameters include the utilization of Maxima H-minus reverse transcriptase, transitioning the RT salt from KCl to NaCl or CsCl, performing RT in the presence of 5% polyethylene glycol (PEG), and incorporating GTPs or dCTPs to enhance and stabilize the TS reaction. A distinctive feature of Smart-seq3 is its integration of full-length transcriptome coverage with a 5' UMI RNA counting strategy, which elevates the precision of transcript counting without sacrificing overall coverage. In this approach, a TSO primer is constructed, comprising a partial Tn5 motif, an 11 base pair tag sequence, an 8 base pair UMI sequence, and three riboguanosines. Following sequencing, the 11 base pair tag is utilized to unequivocally differentiate 5' UMI-labeled reads from internal reads. Within a single sequencing reaction of Smart-seq3, both 5' UMI-labeled reads and internal reads that span the entire transcript without UMIs are collected. This approach allows for the accurate quantification of the original transcripts using UMI reads, correcting for the nonlinear PCR amplification bias. The reconstruction of full-length transcripts is achieved through the use of internal reads. Many current sequencing library construction methods incorporate cell barcodes and UMI tag strategies. However, as these can only be introduced at the ends of the cDNA, the resulting sequenced cDNA sequences are limited to one end of the transcript, leading to the loss of significant sequence information in the middle of the transcript. Consequently, tag-based methods are primarily employed for gene expression quantification and are unsuitable for isoform identification or splicing. Despite their ability to capture full-length transcripts, Smart-seq and Smart-seq2 are unable to utilize barcodes or UMIs to tag transcripts, making them incompatible with high-throughput, parallel single-cell sequencing. Additionally, without a UMI tag, these methods cannot address the amplification bias introduced by PCR. However, Smart-seq3 overcomes the limitation of incompatibility between full-length transcript coverage and UMI by utilizing a special TSO primer.

Smart-seq-total was designed to address the limitation of capturing only poly(A)+ RNA molecules in the previous Smart-seq technologies (Isakova et al., 2021). The primary advancement in this approach lies in the utilization of *Escherichia coli* poly(A) polymerase to add adenine tails to the 3' end of RNA molecules. Consequently, all poly(A)+ RNAs are reverse-transcribed using a poly(T) primer that incorporates a UMI, along with the TSO. This modification enables Smart-seq-total to capture diverse RNA forms concurrently, encompassing protein-coding, long-noncoding, microRNA, and other noncoding RNA transcripts within a single cell. Such an approach facilitates the exploration of regulatory connections between coding

424 and noncoding transcripts in a cell, offering insights into the intricate regulatory
 425 landscape. However, it is important to note that Smart-seq-total does have some
 426 limitations. Firstly, it cannot assess circRNA. Secondly, it results in the loss of the
 427 endogenous polyadenylation status of transcripts. Despite these drawbacks,
 428 Smart-seq-total exhibits significant potential for uncovering noncoding regulatory
 429 patterns governing cellular functions and contributing to the definition of cellular
 430 identity.

431

432 *CEL-seq*

433 Cell Expression by Linear Amplification and Sequencing (CEL-seq) was the
 434 pioneering method to utilize linear strand-specific in IVT for RNA amplification from
 435 single cells (Hashimshony et al., 2012). Initiation of the procedure includes the
 436 synthesis of the first-strand cDNA using a primer featuring an anchored poly(T), a
 437 specific barcode, the 5' Illumina sequencing adaptor, and a T7 promoter.
 438 Subsequently, the second strand is generated to produce double-stranded cDNA
 439 containing a T7 promoter. Combined cDNA samples from several cells undergo IVT,
 440 initiated by the T7 promoter, enabling linear amplification of cDNA. The resulting
 441 amplified RNAs are then converted into cDNAs for sequencing. By employing linear
 442 amplification, CEL-seq minimizes amplification bias, delivering a more sensitive and
 443 reproducible outcome compared to full-length cDNA coverage techniques like
 444 Smart-seq. However, CEL-seq has limitations, such as the inability to detect miRNAs
 445 and other poly(A)- transcripts and difficulty in distinguishing alternative splice forms
 446 due to its strong 3' bias.

447 CEL-seq2, an improved version of CEL-seq increased sensitivity, reduced costs,
 448 and decreased hands-on time (Hashimshony et al., 2016). To mitigate mRNA
 449 molecule counting biases, CEL-seq2 introduces 5 base pair UMIs upstream of the
 450 barcode. Utilizing the SuperScript II Double-Stranded cDNA Synthesis Kit, coupled
 451 with a modification of the CEL-seq primer length, notably enhances the efficiency of
 452 the RT reaction. CEL-seq2 outperforms the original CEL-seq method by detecting
 453 twice as many transcripts and 30% more genes per cell. However, it is important to
 454 note that despite its notable 3' bias, CEL-seq2 does not offer information on the
 455 majority of splicing events. Nevertheless, its increased sensitivity and capability for
 456 individual transcript counting provide a clear advantage for various applications in
 457 transcriptomics.

458

459 *SUPeR-seq*

460 In Single-cell Universal Poly(A)-independent RNA Sequencing (SUPeR-seq), random
 461 primers with fixed anchor sequences are utilized instead of the commonly used
 462 oligo-dT primers in cDNA synthesis. This enables the detection of both poly(A)+ and
 463 poly(A)- RNAs within a single cell (Fan et al., 2015b). The process involves the
 464 utilization of the random primers with a fixed anchor sequence (AnchorX-T15N6) to
 465 reverse transcribe total RNAs into first-strand cDNAs. Following the synthesis of the
 466 initial cDNA strand, ExoSAP-IT is employed to digest excess primers, preventing the
 467 formation of primer-dimer complexes. Using terminal deoxynucleotidyl transferase

468 and dATP with 1% ddATP, a poly(A) tail is appended to the 3' end of the initial
 469 cDNA strand. Subsequently, poly(T) primers with an alternative anchor sequence
 470 (AnchorY-T24) are applied to generate the second-strand cDNA, which undergoes
 471 amplification through PCR using AnchorY-T24 and AnchorX-T15 primers for
 472 subsequent sequencing. SUPeR-seq has been utilized to investigate the regulatory
 473 mechanisms of circRNAs during early embryonic development in mammals.
 474 However, it presents challenges for high-throughput sequencing as well as molecule
 475 counting due to the absence of UMIs and cell barcodes.

476

477 *MATQ-seq*

478 In 2017, Sheng et al. introduced the Multiple Annealing and dC-tailing-based
 479 Quantitative single-cell RNA-seq (MATQ-seq) method (Sheng et al., 2017). This
 480 technique incorporates barcodes and UMIs to sequence both polyA+ and polyA-
 481 RNAs, distinguishing it from SUPeR-seq. The procedure encompasses converting
 482 total RNAs into first-strand cDNA using primers designed for Multiple Annealing and
 483 Looping-Based Amplification Cycles (MALBAC). These primers contain mainly G,
 484 A, and T bases, along with MALBAC-dT primers. After RT, the first-strand cDNA
 485 undergoes dC tailing, followed by the synthesis of second-strand cDNA using
 486 G-enriched MALBAC primers. UMIs are presented during second-strand synthesis.
 487 Unlike Smart-seq2 and SUPeR-seq, MATQ-seq utilizes UMIs to significantly reduce
 488 leading or trailing-end bias in HEK293T transcripts. Additionally, MATQ-seq
 489 exhibits higher sensitivity than Smart-seq2 and SUPeR-seq in capturing polyA- RNAs,
 490 with a capture efficiency of $89.2\% \pm 13.2\%$. This improvement enhances the
 491 efficiency of detecting the low-abundance genes. MATQ-seq's high accuracy and
 492 sensitivity allow for the detection of subtle differences in gene expression among
 493 individual cells within the same population. However, similar to SUPeR-seq, the
 494 time-consuming cell isolation method involving a mouth pipette limits MATQ-seq's
 495 throughput.

496

497 *FLASH-seq*

498 FLASH-seq, a swift and highly profound full-length scRNA-seq method, was
 499 developed by Hahaut et al. in 2022 (Hahaut et al., 2022). Several key modifications
 500 were introduced to enhance the efficiency of the Smart-seq2 protocol: (1) it combined
 501 the reverse transcription and cDNA preamplification, streamlining the process; (2)
 502 Superscript IV, a more processive reverse transcriptase, replaced Superscript II, and
 503 the RT reaction time was shortened; (3) The amount of dCTP was increased to favor
 504 the C-tailing activity of Superscript IV and enhance the template-switching reaction;
 505 (4) The 3'-terminal locked nucleic acid guanine in the template-switching
 506 oligonucleotide was replaced with riboguanosine; (5) The reaction volume was
 507 reduced to 5 μL (Hahaut et al., 2022). These modifications collectively led to a
 508 substantial decrease in both time and cost. FLASH-seq can be completed in
 509 approximately 4.5 hours, making it 2-3.5 hours faster than other methods like
 510 Smart-seq2. The cost per cell is lower than other commercial and non-commercial
 511 methods, comparable to Smart-seq3, and amounts to less than \$1 per cell.

512 Furthermore, FLASH-seq can detect a substantial number of SNPs. It is considered
 513 suitable for researchers seeking affordable, automation-friendly, robust, and efficient
 514 methods for single-cell transcriptional profiling.

515

516 *High-throughput scRNA-seq technologies*

517 *Strategy for developing high-throughput scRNA-seq*

518 In the early stages of scRNA-seq, micromanipulators or LCM were employed to
 519 isolate individual cells for separate transcriptome amplification and library
 520 construction. However, these methods had limitations as they could only analyze a
 521 few cells in a single experiment. The introduction of cell-specific barcodes
 522 revolutionized the field, allowing thousands of single-cell transcriptomes to be mixed
 523 together for library construction and sequencing in a single run, enabling
 524 high-throughput parallel sequencing. A notable advancement in this direction was the
 525 development of the MARS-seq method, which combined FACS with automatic liquid
 526 handling to successfully sequence thousands of cells in a single experiment (Jaitin et
 527 al., 2014). Three levels of barcodes were used to label cells, plates, and mRNAs,
 528 facilitating the mixing of all materials for subsequent automated processing. Similar
 529 to MARS-seq, STRT-Seq-2i aimed to increase sequencing throughput by
 530 implementing a specialized FACS and barcoding protocol (Hochgerner et al., 2017).
 531 This method utilized a custom aluminum plate with 9,600 wells arranged in 96
 532 subarrays of 100 wells each, enabling the simultaneous sequencing of 9,600 cells in
 533 one run. However, despite the improvements in throughput achieved by these
 534 plate-based methods, the number of cells that could be analyzed was still limited.

535 The introduction of microfluidic techniques has fundamentally addressed the
 536 challenges associated with high-throughput single-cell operations. In 2012, the
 537 Fluidigm C1 system became the first commercially available automated microfluidic
 538 platform designed for the automatic isolation of cells, cell lysis, cDNA synthesis,
 539 amplification, and library preparation for 96 single cells simultaneously. However, the
 540 processing capacity of this system was limited and fell short of meeting the demands
 541 for high-throughput parallel sequencing. A revolutionary breakthrough occurred in
 542 2015 with the advent of droplet-based microfluidic scRNA-seq technologies. These
 543 approaches, exemplified by methods such as those developed by Klein et al. (2015)
 544 and Macosko et al. (2015), enabled the concurrent processing of thousands of cells
 545 (Klein et al., 2015; Macosko et al., 2015). This marked a significant advancement,
 546 allowing for truly massive parallel sequencing in the field of single-cell genomics.

547 Subsequently, additional high-throughput parallel sequencing strategies were
 548 developed, including sci-RNA-seq (Cao et al., 2017) and SPLiT-seq (Rosenberg et al.,
 549 2018). These approaches employ a combinatorial indexing method to label cells
 550 without the physical compartmentalization of single cells. Notably, these technologies
 551 exhibit high cell labeling efficiencies, are straightforward to operate, and can
 552 substantially reduce the cost of sequencing. In the following, we will delve into a
 553 detailed discussion of these innovative technologies.

554

555 *Plate-based high-throughput scRNA-seq methods*

556 High-throughput scRNA-seq methods based on microplates involve sorting cells onto
 557 microplates through FACS and utilizing barcodes to label cell transcripts for
 558 subsequent high-throughput sequencing. These methods offer the advantage of
 559 processing any number of individual cells without significantly impacting the cost per
 560 cell. On the other hand, alternative high-throughput approaches, such as droplet-based
 561 technologies like Drop-seq (Macosko et al., 2015), InDrop (Klein et al., 2015), and
 562 10x Chromium (Zheng et al., 2017), tend to be cost-effective primarily when
 563 analyzing a very large number of cells simultaneously. The sequencing throughput of
 564 plate-based methods is constrained by the number of microplates utilized. The
 565 following section will provide a detailed discussion of these technologies.
 566

567 *Quartz-seq*

568 To decipher the biological functions and fundamental causes of non-genetic cellular
 569 heterogeneity, Sasagawa et al. developed the simple yet highly quantitative
 570 Quartz-seq technique (Sasagawa et al., 2013). Quartz-seq enhances the simplicity and
 571 quantitative performance of whole-transcript amplification by addressing three critical
 572 aspects. First, in order to overcome the problem of overabundance of byproducts in
 573 previous poly(A) tail reaction-based whole-transcript amplification techniques,
 574 Quartz-seq combines exonuclease I treatment, a regulated poly(A) tail, and an
 575 optimized suppression PCR. This strategic approach completely eliminates the
 576 synthesis of byproducts, simplifying subsequent scRNA-seq analysis. Secondly, the
 577 technology adopts a robust DNA polymerase (MightyAmp DNA Polymerase)
 578 optimized for a single-tube reaction. This choice of PCR enzyme enhances cDNA
 579 yield, improves the reproducibility of whole-transcript amplification replication, and
 580 reduces the number of required PCR cycles. Finally, Quartz-seq optimizes the
 581 efficiencies of RT and second-strand synthesis by adjusting the annealing temperature.
 582 Notably, all steps of this method are consolidated into a single PCR tube, eliminating
 583 the need for purification and involving only six reaction steps per single cell. This
 584 streamlining significantly simplifies the Quartz-seq approach, facilitating its
 585 high-throughput implementation. Beyond its simplicity, Quartz-seq exhibits high
 586 quantitative accuracy, reproducibility, and sensitivity. Consequently, it can discern
 587 various types of non-genetic cellular heterogeneity and differentiate between distinct
 588 cell types and cell-cycle phases within the same cell type.

589 Quartz-seq2 (Sasagawa et al., 2018), an innovative high-throughput scRNA-seq
 590 method, was developed as an extension of Quartz-seq. This approach involves sorting
 591 single cells using FACS into a 384-well plate, followed by RT using a primer that
 592 combines an oligo-dT sequence, a cell barcode sequence, and a UMI sequence. By
 593 applying advancements in multiple molecular biological stages, including a major
 594 upgrade of poly(A) tagging, quartz-seq2 achieves excellent UMI conversion
 595 efficiency. Notably, Quartz-seq2 utilizes a poly(A) tagging strategy based on the
 596 combination of T55 buffer and the increment temperature condition, resulting in an
 597 approximately 3.6-fold increase in the amount of cDNA.

598 The UMI conversion efficiency of Quartz-seq2 is notably high, ranging from 32
 599 to 35%, surpassing that of other methods such as CEL-seq2, SCRIB-seq, and

600 MARS-seq (7%~22%). Quartz-seq2 can identify more transcripts from fewer
 601 sequence reads at a lower cost because of its increased efficiency. Similar to
 602 MARS-seq, Quartz-seq2 employs FACS for cell sorting, a process that requires
 603 skilled workers. Despite this requirement, the technology's enhanced UMI conversion
 604 efficiency and cost-effectiveness make it a valuable method for scRNA-seq
 605 applications.

606

607 *MARS-seq*

608 MARS-seq, an automated and highly parallel RNA single-cell sequencing technology
 609 developed in 2014 (Jaitin et al., 2014), revolutionized the field by enabling the
 610 counting of unique RNA molecules through the introduction of UMIs in the oligo-dT
 611 primer. Single cells are sorted using FACS into 384-well plates as part of the
 612 MARS-seq procedure. The RT and library construction processes follow the CEL-seq
 613 protocol, ensuring a systematic and reproducible approach. One of the key
 614 innovations of MARS-seq lies in its automation, with every step of the method being
 615 executed by a liquid-handling robot. This automation improves the technique's
 616 repeatability and leads to a significant boost in throughput. The high-throughput and
 617 automated nature of MARS-seq makes it applicable to diverse tissues and organs in
 618 both normal and disease states. By delineating the cell-type and cell-state
 619 compositions of tissues, MARS-seq contributes to a comprehensive understanding of
 620 these biological entities, linking this information to detailed genome-wide
 621 transcriptional profiles.

622 In 2019, the research team developed an integrated pipeline for index sorting and
 623 massively parallel single-cell RNA sequencing (MARS-seq2.0), building on the
 624 foundation of the MARS-seq method (Keren-Shaul et al., 2019). MARS-seq2.0 offers
 625 the capability to efficiently sequence 8,000~10,000 cells in a single run, representing
 626 a significant enhancement in throughput. Notably, the method achieves an eight-fold
 627 reduction in the volume of the RT reaction, decreasing it from 4 μ L to 500 nL. The
 628 cost of preparing single-cell libraries will drop six-fold as a result of this volume
 629 reduction. MARS-seq2.0 is a 3'-based scRNA-seq technique, which limits its use for
 630 determining alternative splicing isoforms or particular sequences at the 5' end of the
 631 gene. This is an essential point to remember. Despite this limitation, the integration of
 632 indexed FACS sorting with scRNA-seq in MARS-seq2.0 proves beneficial for
 633 identifying rare subpopulations and processing rare cells in human clinical samples.
 634 Moreover, MARS-seq2.0 provides a flexible platform that enables simultaneously
 635 obtain multiple layers of information on the same single cell, encompassing genetics,
 636 signaling, epigenetics and spatial location by combining unbiased transcriptional
 637 mapping with large numbers of fluorescent markers. This multifaceted approach
 638 contributes to a deeper molecular understanding of physiological processes and
 639 diseases.

640

641 *SCRB-seq*

642 To economically characterize the major patterns of gene expression variation across
 643 heterogeneous populations, Soumillon et al. developed single-cell RNA barcoding and

sequencing (SCRB-seq) based on the Smart-seq protocol to profile mRNAs from large numbers of cells using minimal reagents and sequencing reads per cell (Soumillon et al., 2014). In SCRB-seq, FACS is used to sort individual cells into a 384-well plate. Poly(A)+ mRNA is converted to cDNA by RT, which uses a template-switching reverse transcriptase and RT primers made of barcodes, UMIs, and a poly(T) primer. After that, strand information is preserved and the decorated cDNA from several cells is combined and amplified for sequencing using a modified transposon-based fragmentation technique that enriches for the 3' end. SCRB-seq technology is capable of sequencing about 12,000 single cells, providing deep and full-length transcriptome coverage sequencing. Moreover, it requires roughly two times fewer enzymatic reactions, purifications, and liquid transfer steps than the MARS-seq approach (Jaitin et al., 2014). SCRB-seq, in contrast to Smart-seq, adds unique cell barcodes during RT, making it easier to identify reads that come from the same cell and boosting sequencing throughput. Nevertheless, sequencing a larger number of single cells still faces challenges.

659

660 *STRT-seq-2i*

661 STRT-seq-2i is a dual-index 5' single-cell and nucleus RNA-seq method designed to
662 significantly increase throughput (Hochgerner et al., 2017). It utilizes a specially
663 designed 9,600-microwell plate, contributing to enhanced efficiency. The microwell
664 array allows for imaging verification of single-cell wells, reducing the occurrence of
665 doublets in a single well. In addition to maintaining several advantages, such as
666 5'-end reads that reveal transcription start sites, the addition of UMIs for absolute
667 quantification, and the use of single-read sequencing rather than paired-end
668 sequencing to maximize cost efficiency, STRT-seq-2i is still compatible with the
669 previously described STRT-seq method. This technique has demonstrated its
670 adaptability to various experimental contexts by being used to examine the
671 transcriptional profile of both fresh single mouse cortical cells and frozen
672 post-mortem human cortical nuclei.

673

674 *Microfluidics-based high-throughput scRNA-seq methods*

675 Droplet- and microwell-based platforms stand out as the predominant technologies for
676 high-throughput scRNA-seq, capable of profiling transcriptomes from approximately
677 10,000 individual cells in a single experiment. Both methods basically involve
678 separating individual cells into many nanoliter-sized containers (such as microwells or
679 water-in-oil droplets) that contain the chemicals required for RT. The integration of
680 cell barcoding strategies into these microfluidic platforms significantly enhances
681 throughput while concurrently reducing costs when compared to both nonmicrofluidic
682 methods and microfluidic methods featuring valves. This advancement is particularly
683 advantageous for biomedical research applications demanding the comprehensive
684 transcriptomic profiling of a vast number of cells.

685

686 *Droplet-based scRNA-seq technologies*

687 *InDrop*

The fundamental technology of inDrop (Klein et al., 2015) consists of encasing single cells into droplets that contain lysis buffer, RT reagents, and barcoded hydrogel microspheres (BHMs). Each BHM contains $\sim 10^9$ photocleavable barcoded primers (147,456 distinct barcodes). The microfluidic device employed in this technology incorporates inlets for carrier oil, cells, lysis/RT reagents, and BHMs, along with an outlet for droplet collection. Each BHM is covalently linked to barcoded primers, featuring a T7 RNA polymerase promoter, an Illumina sequencing adaptor, a unique cell barcode, UMI, and a poly(T) tail, connected via a photo-releasable bond. All BHMs within a sample share the same cell barcode for sample distinction but possess distinct UMIs for precise transcript counting. After encapsulation, ultraviolet (UV) exposure facilitates the photo-release of barcoded primers, allowing mRNA from lysed cells to be barcoded during cDNA synthesis while remaining confined to the same droplet. Subsequently, cDNAs from all cells are pooled post-droplet breakage for library construction and sequencing, following the CEL-seq protocol (Hashimshony et al., 2012). InDrop's intrinsic scalability, unlike conventional methods, is not constrained by the number of reaction chambers. Additionally, the operation processes are simplified by conducting lysis and RT within the droplets. However, a notable drawback of inDrop is its relatively low mRNA capture efficiency (~7%), rendering genes with transcript abundances below 20~50 transcripts challenging to reliably detect in a single cell.

708

709 *Drop-seq*

Similar to inDrop, Drop-seq is designed for the high-throughput analysis of mRNA expression in thousands of single cells. It achieves this by co-encapsulating each cell with a uniquely barcoded bead within nanoliter-scale water-in-oil droplets for simultaneous processing (Macosko et al., 2015). In contrast to inDrop, which utilizes barcoded hydrogel microspheres, Drop-seq employs beads made of an unchanging hard resin. These resin beads are directly synthesized with barcoded primers, which include a poly(T) sequence for mRNA capture, a cell barcode, a UMI, and a universal PCR handle for amplification. Each bead contains over 10^8 different primers. Following cell lysis in the droplet, released mRNAs hybridize with the primers' poly(T) tails on companion beads to generate single-cell transcriptomes that are affixed to microparticles (STAMPs). Following droplet breakup, thousands of STAMPs are pooled together, reverse-transcribed, PCR-amplified, and sequenced in a single reaction. The Drop-seq method boasts a significantly larger number of unique barcodes (16,777,216) compared to inDrop (147,456), enabling cost-effective and rapid high-throughput analysis.

725

726 *10x Genomics*

727 The 10x Genomics system (Zheng et al., 2017) is a fully commercial platform that
728 shares some similarities with inDrop and Drop-seq. The Gel Bead in Emulsion (GEM)
729 is the fundamental component of this method. To create GEM, an 8-channel
730 microfluidic device is used. In about 6 minutes, this chip can produce 100,000 GEMs,
731 each of which can contain thousands of cells. Every gel bead in the GEM contains

732 barcoded oligonucleotides comprising Illumina adapters, 10x barcodes, UMIs, and a
 733 poly(T) tail, which primes poly(A)+ RNA transcripts. After co-encapsulation of cells
 734 and gel beads into droplets, cell lysis occurs immediately, releasing mRNAs.
 735 Following this, gel beads dissolve and release the barcoded oligonucleotides, enabling
 736 RT of poly(A)+ RNAs. The RT reaction takes place within each individual droplet,
 737 resulting in cDNA molecules that possess a shared barcode per GEM, a unique UMI,
 738 and end with a TSO at the 3' end. The barcoded cDNA molecules are then combined
 739 for PCR amplification, adhering to the Smart-seq methodology, after the emulsion has
 740 been broken.

741 InDrop, Drop-seq, and Chromium are three similar platforms that employ
 742 droplet-microfluidic approaches to isolate single cells for high-throughput sequencing.
 743 All three methods can process tens of thousands of cells rapidly each day. The three
 744 technologies differ in the following four aspects:

745 First, inDrop and 10x Genomics use hydrogel microspheres, while the beads
 746 used in Drop-seq are hard resin. Encapsulation of the soft and pliable hydrogel beads
 747 used in inDrop and 10x Genomics, which are densely packed in the microfluidic
 748 channel, can be synchronized to produce a super-Poisson distribution. A double
 749 Poisson distribution governs the encapsulation of tiny hard beads for a single cell and
 750 bead in the same droplet. Thus, inDrop and 10x Genomics can achieve significantly
 751 higher cell capture efficiency compared to Drop-seq. It has been reported that the cell
 752 capture rate of 10x Genomics is about 50% (Zheng et al., 2017).

753 Secondly, inDrop and 10x Genomics use hydrogel beads, allowing the primers to
 754 be fixed inside the beads, whereas the primers of Drop-seq can only be fixed on the
 755 surface of the smaller hard beads. After encapsulation, inDrop uses
 756 UV-irradiation-induced cleavage to release primers. The 10x Genomics releases all of
 757 the primers into the solution by directly dissolving the beads to improve the capture
 758 efficiency of mRNA. In contrast, the primers of Drop-seq cannot be released from the
 759 beads, and mRNA molecules hybridize to the poly(T) tail on the beads to form
 760 STAMPs for RT, which could be a drawback of Drop-seq compared to inDrop and
 761 10x Genomics.

762 Thirdly, in Drop-seq, droplets are split immediately as soon as the mRNA and
 763 primer hybridize, and all STAMPs are mixed together to perform RT. Instead, in the
 764 inDrop and 10x Genomics methods, RT reagents are co-encapsulated into droplets,
 765 and RT reactions are conducted independently within each droplet, which is beneficial
 766 for improving the specificity of cDNA conversion, enhancing relative yield, and
 767 reducing reagent consumption (Streets et al., 2014).

768 Finally, the three platforms adopt different library construction strategies. While
 769 Drop-seq and 10x Genomics use a template-switching procedure akin to the
 770 well-known Smart-seq chemistry, inDrop uses the CEL-Seq approach. They so inherit
 771 the benefits and drawbacks of Smart-seq and CEL-seq, respectively.

772

773 *MULTI-seq*

774 To deliver cell-specific barcodes during RT, all of the droplet-based techniques
 775 previously discussed generally use the co-encapsulation strategy, which entails

776 simultaneously encapsulating cells and barcoded beads. The MULTI-seq method
777 (sample multiplexing for single-cell and single-nucleus RNA sequencing using
778 lipid-tagged indices) was recently introduced by McGinnis et al. as an alternative
779 strategy that makes unique use of cell barcodes (McGinnis et al., 2019b). In this
780 technology, DNA barcodes are labeled onto the plasma membranes of single cells by
781 hybridization to an ‘anchor’ lipid-modified oligonucleotide (LMO). The hydrophobic
782 5’ lignoceric acid amide of the ‘anchor’ LMO binds to membranes; hybridization to a
783 ‘co-anchor’ LMO with a 3’ palmitic acid amide amplifies the hydrophobicity of the
784 complex, extending membrane retention. The 3’ poly(A) capture sequence, the 8 bp
785 sample barcode, and the 5’ PCR handle make up the LMO. Each single cell or nuclei
786 carried by the LMOs is co-encapsulated with an mRNA capture bead into an emulsion
787 droplet generated by the 10x Genomics system. Sample demultiplexing is made
788 possible by the release of endogenous mRNAs and LMOs upon in-drop cell lysis,
789 which both hybridize to the mRNA capture bead and attach to a common cell barcode
790 during RT. Endogenous mRNAs and LMOs are separated by size selection following
791 amplification, enabling pooled sequencing at user-defined ratios. Any cell type or
792 nucleus from any species with an accessible plasma membrane can be barcoded using
793 this method. Moreover, this approach involves minimal sample handling, preserving
794 cell viability and endogenous gene expression patterns.

795

796 *Microwell-based scRNA-seq Methods*

797 *CytoSeq*

798 CytoSeq, a highly scalable scRNA-seq method, can simultaneously analyze a few
799 thousand cells and can be easily scaled to 10,000 or 100,000 cells, with the detection
800 of approximately 100 genes per cell (Fan et al., 2015a). This method employs a
801 recursive Poisson strategy to adjust the number of cells in suspension, facilitating
802 high-throughput cell settling in 1/10 of 100,000 microwells by gravity. Due to the
803 nanoliter volumes used in the reactions, the cost of library preparation is exceptionally
804 low. Additionally, CytoSeq is advantageous as it is not restricted to specific cell sizes
805 and shapes, enabling the study of expression profiles in large and heterogeneous cell
806 populations. This flexibility allows for the detection of rare cell types within a large
807 background population.

808

809 *Seq-Well*

810 Seq-Well is a portable, cost-effective, user-friendly, and efficient scRNA-seq method
811 designed for low-input samples (Gierahn et al., 2017). This approach utilizes a
812 picowell array where barcoded mRNA capture beads and cells are loaded, with each
813 well accommodating one cell and nearly one bead. After the cells settle into the wells
814 by gravity, a semipermeable membrane seals the array, creating a unique environment
815 for each well that allows buffer exchange but prevents the migration of
816 macromolecules. Subsequently, cells are lysed, and the process of amplification and
817 sequencing is carried out. With approximately 86,000 subnanoliter wells, Seq-Well
818 plates enable the simultaneous analysis of transcriptomes in thousands of cells from
819 diverse sources. This method is particularly well-suited for low-sample inputs, such as

tissue pinches, fine-needle aspirates, and challenging-to-study cells like hepatocytes and granulocytes (Kumar, 2021). Notably, Seq-Well's implementation requires only a picowell array, a pipette, a polycarbonate membrane, an oven or heat source, a clamp, and a tube rotator to generate a stable cDNA product. This simplicity makes it adaptable to resource-limited environments such as clinic and remote locations (Aicher et al., 2019).

826

827 ICELL8

ICELL8 is a microwell-based microfluidic system that enhances throughput by incorporating a microchip containing 5,184 nanowells, allowing for the capture and processing of approximately 1,300 single cells (Goldstein et al., 2017). Each nanowell contains preprinted oligonucleotides, which include an oligo-(dT₃₀) primer, a well-specific barcode sequence, and a UMI. The process involves dispensing single-cell suspensions into the microchip nanowells using the multi-sample nanodispenser (MSND) and subsequently lysing the cells through freeze-thaw cycles. Following cell lysis, RT is carried out to synthesize cDNAs, employing the SCRB-seq method. Ultimately, cDNAs from hundreds of cells are pooled into a single tube for library construction. The method offers several advantages: (i) the use of the MSND instrument for accurate dispensing of cells into nanowells, eliminating errors associated with manual pipetting; (ii) incorporation of imaging software to identify nanowells containing viable single cells, ensuring that only wells with single cells are processed into sequencing libraries, resulting in a low cell multiplet rate (<3%); (iii) the capability to load up to eight experimental conditions across one array using a multi-sample nanodispenser, enabling the simultaneous processing of 800 to 1,400 cells on one chip with 5,184 nanowells in a single experiment.

To enhance the capture rate of ICELL8 and enable the simultaneous processing of more than 5,000 cells for sequencing libraries, Orr Shomroni et al. introduced a novel scRNA-seq method called CellenONE-ICELL8. This method combines the ICELL8 processing instrument with the CellenONE isolation and sorting system (Shomroni et al., 2022). CellenONE relies on image-based single-cell isolation, enabling the selection of highly purified individual cells based on parameters such as cell morphology, size, or fluorescence markers before subsequent sample processing and sequencing. The integration of both instruments in CellenONE-ICELL8 significantly improves cell capture efficiency compared to the ICELL8 system alone, raising the number of captured cells from the typical 1,200 to 1,400 cells to over 3,300 cells. Furthermore, the utilization of full-length chemistry (SMART-seq technology) in this method can detect non-coding RNAs, especially lengthy intergenic non-coding RNAs, as well as intronic and intergenic sequences.

858

859 Combinatorial indexing-based high-throughput scRNA-seq technologies

860 Sci-RNA-seq

861 Cao et al. developed the first combinatorial indexing method for high-throughput
 862 scRNA-seq, termed sci-RNA-seq (Cao et al., 2017). This innovative method
 863 facilitates the analysis of the transcriptomes of large numbers of single cells or nuclei,

864 providing 3' coverage and high-depth sequencing through the use of double UMI
 865 barcoding. The scalability of sci-RNA-seq allows for the generation of approximately
 866 4×10^4 individual cell transcriptomes in a single experiment when employing
 867 indexing up to 576×960 . This scalability enables the processing of more cells with
 868 sublinear cost scaling by incorporating additional rounds of indexing or/and using
 869 more barcoded RT and PCR primers. But it's crucial to remember that sci-RNA-seq is
 870 not without its drawbacks. These include laborious experimental procedures, the high
 871 expense of high-throughput transposition reactions, and a notable cell loss brought on
 872 by FACS sorting.

873

874 *SPLiT-seq*

875 SPLiT-seq represents another innovative combinatorial indexing method designed for
 876 scRNA-seq analysis, allowing the examination of fixed tissues preserved in 1.33%
 877 formaldehyde (Rosenberg et al., 2018). Unlike sci-RNA-seq, SPLiT-seq inserts
 878 second and third-round barcodes into cDNA by ligation. This approach offers a
 879 simpler, gentler, and more cost-effective workflow. The first-round barcodes of
 880 SPLiT-seq can act as sample identifiers, thus enabling highly multiplexed parallel
 881 sample processing. SPLiT-Seq is especially well suited for the analysis of fixed,
 882 difficult-to-completely disaggregate cells or nuclei produced from clinically relevant
 883 tissue samples. However, some limitations should be considered, including potential
 884 chemical modifications to mRNA caused by aldehyde-based cell fixation, suboptimal
 885 reaction efficiency of RT and ligation within fixed cells due to the intricate
 886 cross-linked intracellular environment, and potential degradation of RNA quality
 887 during the fixation process, leading to reduced detected gene levels.

888

889 *Single nuclei RNA-sequencing (snRNA-seq)*

890 *Overview of snRNA-seq*

891 ScRNA-seq is a potent tool for exploring cell types, functional processes, and
 892 dynamic states in intricate tissues. However, its application is limited when dealing
 893 with archived samples or tissues that cannot be easily dissociated, preventing the
 894 exploration of new cell types or crucial information related to immunity and disease.
 895 To address this difficulty, scientists have resorted to single-nuclei RNA sequencing
 896 (snRNA-seq) technologies, in which RNA sequencing experiments are performed
 897 using nuclei as proxies rather than entire cells. Several snRNA-seq methods have
 898 been developed to analyze RNA in single nuclei obtained from frozen, lightly fixed,
 899 or fresh tissues, including DroNc-Seq (Habib et al., 2017), Div-seq (Habib et al.,
 900 2016), snDrop-seq (Lake et al., 2019), etc. These approaches extend the applicability
 901 of transcriptomic analyses to a broader range of sample types and conditions.

902 The use of snRNA-seq techniques has proven valuable in various research areas
 903 due to the high correlation observed between genes detected by snRNA-seq and
 904 traditional scRNA-seq methods (Fischer and Ayers, 2021). These techniques find
 905 application in diverse sample types, such as fresh tissues like the brain (Affinati et al.,
 906 2021), heart (Nicin et al., 2021), kidney (Barwinska et al., 2021; Muto et al., 2021),
 907 pancreas (Basile et al., 2021), muscle (Petrany et al., 2020) or adipose tissue (Sun et

al., 2020b), archived tissues (Basile et al., 2021), plant cells (Conde et al., 2021). Despite these advantages, isolated nuclei are often more adhesive compared to isolated cell types. Therefore, precautions should be taken to prevent clumping, which can lead to inflated doublet rates. Furthermore, it's important to note that certain gene transcripts may exhibit enrichment differences between snRNA-seq and scRNA-seq datasets. For instance, long non-coding RNAs (lncRNAs) are enriched in snRNA-seq datasets, while mitochondrial transcripts, residing in the cytosol, are only present in scRNA-seq datasets (Fischer and Ayers, 2021). Researchers need to carefully evaluate whether such sublocalized genes are crucial for their specific research project before opting for individual nuclei for sequencing.

918

919 *Computational methods for scRNA-seq data*

920 The original data format of scRNA-seq and most of the current scRNA-seq analysis
 921 processes are based on FASTQ files (or compressed format fq.gz). Illumina platform
 922 sequencing data generates BCL format files by default, which can be converted
 923 through CellRanger mkfastq. The analysis process of scRNA-seq include data
 924 preprocessing, processing and extended downstream analysis (Figure 4), among
 925 which data preprocessing includes quality control, read alignment and expression
 926 quantification; data processing includes normalization, batch effect correction,
 927 imputation, feature selection (HVG selection), dimension reduction and clustering,
 928 cell typing annotation, differential expression analysis (DEGs), visualization;
 929 extended downstream analysis includes pseudotime, cell-cell interaction (CCI),
 930 pathway enrichment analysis, gene regulatory network (GRN) and other downstream
 931 analysis. On the whole, scRNA-seq analysis methods have mushroomed emerge in
 932 endlessly, there is no absolutely perfect method that applies to all scenes, it is
 933 important to obtain the biological information from analysis tools and the difficulty is
 934 selecting the most appropriate method. Here, we proposed to summarize common
 935 single-cell transcriptome analysis methods with their advantages and disadvantages as
 936 well as the scope of application to make suggestions.

937

938 *Data pre-processing*

939 The original sequencing data by filtering out low-quality reads and environmental
 940 interference were aligned and quantified with reference genomes. Consequently, the
 941 feature count matrix for each cell and auxiliary files recording other information were
 942 obtained, which were used for downstream data analysis (Figure 4A).

943

944 *Quality control*

945 Low-quality sequencing data was inevitably produced due to the sequencing
 946 instrument problem, artificial operation, cell spontaneous cases, or the existence of
 947 empty droplets, doublets, dead cells, etc. (Chen et al., 2019a; Hao et al., 2021b).
 948 Empty droplets usually appear when the droplet captures extracellular background
 949 transcripts instead of cells (Kolodziejczyk et al., 2015; Ilicic et al., 2016). A highly
 950 subjective method is to determine a UMI threshold according to the knee point and
 951 filter out cells with low UMI count. DropEst (Petukhov et al., 2018), EmptyDrops

(Lun et al., 2019), and DIEM (Alvarez et al., 2020) were then used to enhance the filtering effect. DropletQC (Muskoovic and Powell, 2021) quantifies the nuclear fraction score of unspliced pre-mRNA content. The choice of the MT gene threshold requires a comprehensive consideration of cell physiology factors though it is a dead cell measurement (Subramanian et al., 2022). In recent years, deep-learning-based methods, such as neural-network-based EmptyNN (Yan et al., 2021), and deep-generation-models-based CellBender (Fleming et al., 2019), have also emerged and enabled the effective identification of the background transcripts in empty droplets.

Doublet is the case that two cells are contained in one single drop, which can be divided into homo-doublet and hetero-doublet based on the transcriptional distribution, both obeying Poisson statistics (Bloom, 2018). The vast majority of methods are based on gene expression calculations, using prior knowledge or deep learning to obtain the differences between unimodal and bimodal cells, and then train the classifier for screening, e.g., nearest-neighbor-based DoubletFinder (McGinnis et al., 2019a), Scrublet (Wolock et al., 2019); deconvolution-based DoubletDecon (DePasquale et al., 2019), variational-autoencoder-based Solo (Bernstein et al., 2020), and ensemble-algorithm-based Chord (Xiong et al., 2021a). Besides, Scds is another screening method relying on the co-expression-based doublet scoring and binary-classification-based doublet scoring strategy to achieve doublet separation over the scRNA-seq expression data (Bais and Kostka, 2020). A few methods use other features, such as the demuxlet which uses natural genetic variation information guidance experiments and filters computationally (Kang et al., 2018).

Reasonable quality control needs to comprehensively consider both technical and biological factors, which is also the main direction of the current research. A biological data-driven self-learning unsupervised quality control method called ddqc was recently proposed to determine specific thresholds of various GC metrics (Macnair and Robinson, 2023).

980

981 *Read alignment and expression quantification*

982 The remaining high-quality cells after quality control require mapping these short
 983 reads to a specific reference genome for alignment to make the quantification of gene
 984 expression levels. RNA read alignment is usually divided into two steps: alignment of
 985 reads for indexing and mapping RNA splicing sequence, the former step was shared
 986 with DNA read alignment, solving the mismatch problem and setting up index
 987 references; the latter step is unique for RNA read alignment and provides connectivity
 988 information.

989 The early second-generation sequencing results were dozens of pair length base
 990 reads. Seed-to-extend methods (Buhler, 2001) (including MAQ (Li et al., 2008a),
 991 SOAP (Li et al., 2008b), CloudBurst (Schatz, 2009), ZOOM (Lin et al., 2008)),
 992 Burrows-Wheeler Transforming methods (Burrows, 1994) (including SOAP2 (Li et
 993 al., 2009), Bowtie (Langmead et al., 2009), BWA (Li and Durbin, 2009)),
 994 Needleman-Wunsch method (including Novocraft (Hercus, 2009)), and suffix-tree
 995 algorithm method (including MUMmer 2 (Delcher et al., 2002)) are effective tools for

996 reads alignment of million level short-chain DNA sequencing. For Bowtie, an
 997 FM-index method dependent on Burrows-Wheeler Transforming is used, the result
 998 only reports one if the reads have multiple accurate matches, it greatly optimizes the
 999 running memory and alignment speed compared to MAQ (Ferragina and Manzini,
 1000 2001). BWA is another BWT-based alignment method, using the new SAM (Sequence
 1001 Alignment/Map) format to output the alignment results. Based on MAQ and Bowtie
 1002 two short-chain DNA alignment algorithms, Cole Trapnell proposed TopHat, the first
 1003 RNA-seq alignment method for NGS data in 2009 using the 2-bit-per-base encoding
 1004 to achieve efficient alignment of reads to splice sites in the mammalian genome
 1005 without any prior knowledge of the splice sites (Trapnell et al., 2009).

1006 The above methods decrease the alignment accuracy precipitously when the base
 1007 pair length exceeds 50 bp (Gupta et al., 2018; Lebrigand et al., 2020). Two main
 1008 categories are used in NGS single-cell sequencing analysis: Bowtie2-based methods
 1009 and seed-strategy-based methods (Langmead and Salzberg, 2012). Bowtie2 is an
 1010 upgrade to the Bowtie, retaining the FM-index dependent BWT algorithm core, which
 1011 permits gapped alignment and uses single-instruction multiple-data (SIMD) to extend
 1012 to the long sequencing alignment while increasing the running speed. Based on
 1013 Bowtie2, Daehwan Kim propose TopHat2 (Kim et al., 2013) and HISAT (Kim et al.,
 1014 2015) successively. The main methods for seed strategy are STAR (Dobin et al., 2013)
 1015 and Subread (Liao et al., 2013). Based on the Maximal Mappable Prefix (MMP) ideas,
 1016 STAR adopted the strategy of sequential retrieval to set the longest partial reads
 1017 matching with the reference as seed 1, the rest read will continue to match, in turn is
 1018 called from seed 2 to seed n. It is worth noting that Rsubread implements the first read
 1019 alignment and gene quantification process based entirely on the R language platform
 1020 (Liao et al., 2019).

1021 The gene expression quantification can be divided into pseudo-alignment
 1022 quantification and read-alignment-based quantification. Pseudo-alignment refers to
 1023 the alignment of all reads mapping to the reference genome without the rigorous
 1024 two-step method described above, including the selected k-mers alignment method
 1025 (Sailfish (Patro et al., 2014), Kallisto (Bray et al., 2016), Salmon (Patro et al., 2017),
 1026 RapMap (Srivastava et al., 2016)) and Barcode-UMI-Set (BUS) alignment method
 1027 BUStools (Melsted et al., 2019). Kallisto-BUStools is the latest workflow that uses
 1028 the BUS file format for initial data pre-processing, like the BUStools, the
 1029 pseudo-alignment result and quantification counts are saved in the BUS files (Melsted
 1030 et al., 2021). On the other hand, read-alignment-based methods rely on the result of
 1031 the RNA read alignment method to quantify the gene. CellRanger is the official
 1032 open-source data pre-processing software designated by 10x Genomic company to
 1033 replace Longranger (Zheng et al., 2017). STARSolo is a tool to replace the
 1034 mapping/quantification function for Cellranger, it can realize the analysis of
 1035 multi-platform sequencing data and the quantification of transcriptome features
 1036 beyond gene expression (Kaminow et al., 2021). Other read-alignment-based gene
 1037 expression quantification like UMItools (Smith et al., 2017), zUMIs (Parekh et al.,
 1038 2018), Alevin-fry (He et al., 2022), DropEst (Petukhov et al., 2018), RainDrop
 1039 (Niebler et al., 2020), baredSC (Lopez-Delisle and Delisle, 2022), BCseq (Chen and

1040 Zheng, 2018) use various quality filter and barcode/UMI treatment strategy to
 1041 improve the performance of CellRanger to some extent.

1042 Both CellRanger and STARSolo have a good running speed while processing all
 1043 kinds of single-cell transcriptome datasets, including 10x Chromium, with extremely
 1044 high accuracy. However, under the premise of obtaining almost identical results, the
 1045 latter increased the running speed by at least five times compared to the former, which
 1046 also verifies the purpose of using STARSolo to replace CellRanger by Alexander
 1047 Dobin et al. (Chen et al., 2021a; You et al., 2021; Brüning et al., 2022).

1048

1049 *Data processing*

1050 After making the necessary adjustments to the expression matrix (Normalization,
 1051 Batch Effect Correction, Imputation), biological information can be fully mined from
 1052 single-cell transcriptomic data for analysis. Seurat and Scanpy perform the modular
 1053 and scalable processing of the above processes based on R and Python respectively
 1054 and are currently the mainstream analysis process of single-cell transcriptomic data
 1055 (Satija et al., 2015; Wolf et al., 2018). The conventional analysis process and expected
 1056 processing results can be found in total analytical framework (Figure 4B-D).

1057

1058 *Normalization*

1059 In sequencing processing, due to technical reasons or biological differences between
 1060 cells themselves, may cause library size differences in the same samples (between
 1061 cells) or between different samples (Marinov et al., 2014). The infinite number
 1062 methods process according to the library size, according to the specific principle, they
 1063 can be roughly divided into global-scaling-based normalization, spike-in
 1064 normalization, and other data transformation model normalization.

1065 The global scaling method was originally developed for the bulk RNA analysis
 1066 by scaling the global data with a specific scaling factor (Finak et al., 2015). Counts
 1067 per ten-thousand (CPT) transformation and Count Per Million (CPM) transformation
 1068 are common linear scaling methods, without considering the spike-in count, they scale
 1069 all per UMI/total UMI count equidistantly. Other normalization methods include
 1070 reads per million (RPM) (Mortazavi et al., 2008), trimmed mean of M values (TMM),
 1071 DESeq (Robinson and Oshlack, 2010), upper-quartile scaling (Bullard et al., 2010),
 1072 FPKM (Trapnell et al., 2010), RPKM (Tu et al., 2012) have better stability for
 1073 extreme values compared to linear scaling, therefore have a wider range of
 1074 applications like the RPKM/FPKM. However, when using such methods alone for the
 1075 normalization of the single-cell transcriptome, because of the sparsity and inflated
 1076 false positives of the data, the effect is not acceptable (Evans et al., 2018).
 1077 Improvements are often needed when combined with specific methods. SCnorm uses
 1078 a quantile regression method to evaluate the scale factors among different sequencing
 1079 depth dependence cell groups (Bacher et al., 2017). Based on the assumption of
 1080 Negative Binomial (NB) distribution for gene original count with the true count,
 1081 bayNorm uses an integrated Bayesian model for scRNA-seq data normalization (Tang
 1082 et al., 2020).

1083 Spike-in normalization method can be regarded as another expansion of the

global scaling method, as the scaling factor is calculated from spike-in genes. It is noted that adding information about RNA spike-ins to other methods can also improve the effect of standardization like SCnorm. GRM is a method based on the gamma distribution of the spike-in ERCC molecule concentration in which ERCC is a calibration material commonly used in sequencing (Ding et al., 2015). BASiCS is an automated Bayesian normalization method applying the Poisson hierarchical model to spike-in (technical) genes for cell specific normalization constants inference (Vallejos et al., 2015).

In the above methods, genes were scaled under the assumption of constant intracellular RNA number which can be deceitful, so other transformation models adopt different strategies. Due to the problem of zero-inflated in the single-cell transcriptomic data, some of the models were designed for this purpose, for example, the relative log expression (RLE) method ascend (Senabouth et al., 2019) and the NB-based models like Dino (Brown et al., 2021), scTransform (Hafemeister and Satija, 2019). Other transformation model normalization method like MUREN uses the least trimmed squares (LTS) regression algorithm (Feng and Li, 2021); Sanity uses the log transcription quotients (LTQs) inferred from UMI count as the input of a Bayesian framework to avoid the Poisson fluctuations, as the LTQs vector changes estimate the gene expression values (Breda et al., 2021); PsiNorm is an unsupervised Pareto distribution scale parameter based method to make improvements to the normalization efficiency and accuracy (Borella et al., 2021). Charles Wang made a comparison of total of eight normalization methods including the sctransform, TMM, and DESeq, wherein the sctransform and logCPM (the built-in processing method of Seurat) are least affected by data and are most stable over variable datasets (Chen et al., 2021a).

1109

1110 *Batch effect correction*

Due to the experimental design, sequencing platform, sequencing time, and personnel operation process, different single-cell transcriptome sequencing data will differ significantly in mRNA capture efficiency, and sequencing depth to generate the batch effect among samples (Tung et al., 2017; Hwang et al., 2018; Chen et al., 2019a). Theoretically, the technical variation can be eliminated through experimental strategies, but due to the objective limitations of the experimental process and sequencing instrument errors, the batch effect will inevitably be introduced more or less. Correction with computational methods is necessary to solve imperfect experimental design, usually used methods can be divided into MNN method, latent-space-based method, graph-based method, DL method, and other methods.

Mutual nearest neighbor (MNN) first identifies the most similar cells of the same cell type between different batches, and then uses these cells for batch effect correction, including batchelor (Haghverdi et al., 2018), Scanorama (Hie et al., 2019), Canek (Loza et al., 2022). Another class of methods using MNN is based on latent space after dimension reduction, like Seurat (Satija et al., 2015), BEER (Zhang et al., 2019b), SMNN (Yang et al., 2021a), iSMNN (Yang et al., 2021b). For example, Seurat uses the MNN pairs (called “anchors”) in the canonical correlation analysis

1128 (CCA) latent space to match similar cells while BEER uses the PCA sub-spaces for
 1129 screening poor similar subgroups. SMNN and iSMNN adopt supervised machine
 1130 learning and iterative supervised machine learning separately to refine MN-pairs
 1131 trained from information on pre-correction cell clustering or iterative cell clustering.

1132 Latent space-based methods refer to the method of performing batch effect
 1133 correction in the hidden space or embedding after dimensionality reduction, besides
 1134 the MNN cluster-based strategy, they also contain PCA-related space method
 1135 harmony (Korsunsky et al., 2019), FIRM (Ming et al., 2022), Monet (Wagner, 2020);
 1136 t-SNE space method sc_tSNE (Aliverti et al., 2020) and ZINB-WaVE (Gao et al.,
 1137 2019). Harmony is widely used to remove batch effects between samples, sequenced
 1138 cells are fed into a single common embedding using the PCA (Principal Components
 1139 Analysis) method, then circulated iteratively between maximum diversity clustering
 1140 and linear batch correction until a specific correction factor is assigned to each cell
 1141 which can be used for subsequent batch effect removal. Sc_tSNE method introduces
 1142 gradient descent's algorithm for the traditional t-SNE algorithm optimization, a linear
 1143 correction is used subsequently (Aliverti et al., 2021). ZINB-WaVE was originally
 1144 designed to perform gene extraction in single-cell data, Davide Risso extended this
 1145 method to mini-batch optimization (Risso et al., 2018).

1146 Graph-based methods use cell-gene expression matrix to transform the digital
 1147 information into the spatial constructed graph, where nodes represent different types
 1148 of batches and weights of edges are based on different calculation methods. BBKNN
 1149 uses the k-nearest neighbor cells to construct a graph (KNN graph), and the batch
 1150 effect correction is implemented by merging the graph of individual cells across
 1151 different data sets using the UMAP method, which is also the default method in
 1152 Scanpy workflow (Wolf et al., 2018; Polański et al., 2020). Bo Wang proposed “ghost
 1153 cell” (k-means algorithm cluster center by default) in OCAT to make a bipartite graph
 1154 for cell connection (Wang et al., 2022a).

1155 In recent years, the rapid development of deep learning methods has also
 1156 provided new ideas for batch-effect correction, realizing efficient and
 1157 large-throughput data processing, like INSCT (Simon et al., 2021) (triplet neural
 1158 networks), CLEAR (Han et al., 2022) (Self-supervised contrastive learning),
 1159 BERMUDA (Wang et al., 2019e) (transfer learning), iMAP (Wang et al., 2021a)
 1160 (VAE-GAN), ResPAN (Wang et al., 2022e) (Wasserstein GAN). Some new methods
 1161 are shown to have better results in batch effect correction, for example, based on
 1162 biological prior knowledge from the annotated datasets learned by SciBet, SSBER can
 1163 remove batch effect in a large RNA sequencing data set (Zhang and Wang, 2021). It is
 1164 suggested that before the integration of single-cell transcriptomic data, multiple
 1165 methods should be tested first based on the actual situation of the data, and then the
 1166 most appropriate batch effect removal method should be selected. For example,
 1167 Jinmiao Chen et al and Charles Wang et al conducted a benchmark in 2020 and 2021
 1168 for most of the first three methods mentioned in this review 2.2, respectively, they
 1169 proved that Harmony and Seurat V3 achieve good batch effect correction results in
 1170 most cases, which is in line with the fact that these two methods are still widely used
 1171 today, but there is still a lack of good indicators for deep learning methods (Tran et al.,

1172 2020; Chen et al., 2021a).

1173

1174 *Imputation*

1175 Large numbers of 0 values will be introduced during sequencing (probably over >90%
 1176 zero values in the high-throughput large-scale 10x Genomic sequencing data) (Stegle
 1177 et al., 2015; Talwar et al., 2018). It will interfere with the analysis of downstream
 1178 biological differences, and therefore, the missing data values in the original gene
 1179 expression matrix must be conducted imputation, while effectively distinguishing
 1180 between the technical noise null value and the biological null value.

1181 The gene/cell separated method is mostly applied in the early imputation which
 1182 considers separately the cell similarity (MAGIC (van Dijk et al., 2018), Selmpute (Li
 1183 and Li, 2018), VIPER (Chen and Zhou, 2018), RESCUE (Tracy et al., 2019), scRMD
 1184 (Chen et al., 2020a), scRoc (Ran et al., 2020)) or gene-to-gene relationship (SAVER
 1185 (Huang et al., 2018a), SAVER-X (Wang et al., 2019a), G253 (Wu et al., 2021e), DCA
 1186 (Eraslan et al., 2019), DeepImpute (Arisdakessian et al., 2019)). Overall, these
 1187 methods lack consideration for the data set as a whole and can easily lead to excessive
 1188 imputation or introducing errors (Zhang et al., 2019d). The comprehensive method
 1189 comprehensively considers the connection of cells and genes with each other:
 1190 CMF-Impute and netNMF-sc are the earliest methods to effectively utilize the
 1191 association between cells and genes for imputation (Elyanow et al., 2020; Xu et al.,
 1192 2020a). scIGANs processes the gene expression matrix by a specific GAN model,
 1193 using generated cells training GANs model to impute the dropout (Xu et al., 2020b).
 1194 In recent years, new methods are still being proposed to better solve the impact of
 1195 technical noise on the data outside of dropout and to achieve a better differentiation of
 1196 biological zero values. AutoClass (Li et al., 2022c) achieves processing without
 1197 supervision, while the ALRA method mainly aims at the biological zero values
 1198 (Linderman et al., 2022). scMOO makes a fundamental change to use the latent
 1199 structure of the data to learn deep associations in cell similarity vertical structure and
 1200 total low-rank structure, thus achieving a better imputation effect than a single gene
 1201 expression matrix as an input, but it also puts more memory requirements (Jin et al.,
 1202 2022a). sc-PHENIX utilizes the PCA-UMAP initialization method to achieve a
 1203 nonlinear interpolation of the gene expression (Padron-Manrique et al., 2022). At
 1204 present, there is no definite conclusion on which imputation can achieve the best
 1205 effect. Due to the data set itself, the purpose of downstream analysis will have
 1206 different choices, but there is no doubt that the best imputation method will be able to
 1207 effectively distinguish between technical noise zero value and biological zero value
 1208 with lower calculation requirements (Jiang et al., 2022a; Wen et al., 2022).

1209

1210 *Feature selection*

1211 In order to reduce data dimension to enhance computational analysis efficiency,
 1212 reduce technical noise interference and the risk of model over-fitting, we often choose
 1213 feature selection strategy to select highly variable genes in different cells, instead
 1214 whole data set genes as subsequent analysis, such as cluster (Brennecke et al., 2013;
 1215 Svensson et al., 2017; Jackson and Vogel, 2022).

1216 In bulk RNA-seq analysis, methods for finding differential genes generally
 1217 include fold change (FC) based method, statistical-tests-based method, and
 1218 FC-statistical tests method, the last one has the best screening results and credibility
 1219 obviously (Chung and Storey, 2015).

1220 Early single-cell feature selection approaches lack a correction between the mean
 1221 expression and variances resulting in an excessive proportion of the highly expressed
 1222 gene in the results (Brennecke et al., 2013). EDGE uses an ensemble learning method
 1223 of massive weak learners to learn inter-cellular similarity probabilities, significant
 1224 contributions based on information entropy are extracted as the highly variable genes
 1225 (Sun et al., 2020c). Similarly, SAIC achieves an optimal cell cluster separation based
 1226 on Iterative Clustering final output (Yang et al., 2017). Recently, some new feature
 1227 extraction strategies have been proposed and proved for their stability and
 1228 effectiveness, but the authoritative verification of the performance between them is
 1229 still lacking: including gene expression distribution matrix-based method SCMER
 1230 (Liang et al., 2021b), RgCop (Lall et al., 2021), scPNMF (Song et al., 2021a), SIEVE
 1231 (Zhang et al., 2021e); entropy-based method IEntropy (Li et al., 2022f), infohet
 1232 (Casey et al., 2023); comprehensively considered cluster-based method Triku (A et al.,
 1233 2022), FEAST (Su et al., 2021), etc. Since the vast majority of the above methods
 1234 ignore the integrity of gene dependency, comprehensive methods are proposed, such
 1235 as Triku using a k-nearest neighbor graph method to comprehensively explore and
 1236 classify gene expression patterns, achieve screening for more biologically meaningful
 1237 feature genes without bias; FEAST ranks the feature by f-test on consensus cluster
 1238 and extracts HVG based on feature evaluation algorithm (Wang et al., 2022c).

1239 A few other methods use features other than highly variable genes to represent
 1240 the data set, for example, the scVEGs and scSensitiveGeneDefine methods, using
 1241 high coefficients of variation (CV) as a feature extraction; the BASiCS method
 1242 utilizes the information of spike-in genes (Chen et al., 2016b; Chen et al., 2021b).
 1243 Overall, based on the perspective of accuracy, and biological interpretability, the main
 1244 goal of the current feature selections is to effectively extract the HVG for an effective
 1245 downstream analysis of high-dimensional transcriptome data.

1246

1247 *Dimension reduction*

1248 As the single-cell transcriptome typically includes tens of thousands or more genes, it
 1249 is not conducive to extracting effective information directly. In the actual analysis
 1250 process, we usually need to reduce the dimensionality of the original sequencing data.
 1251 Besides processing the high-dimensional single-cell transcriptome sequencing data
 1252 using the feature selection method mentioned in section 2.4, dimension reduction is
 1253 also an effective method, which can be classified as linear dimension reduction (latent
 1254 dirichlet allocation (LDA)-based method, principal component analysis (PCA)-based
 1255 method) and nonlinear dimension reduction (t-distributed stochastic neighbor
 1256 embedding (t-SNE)-based method, uniform manifold approximation and projection
 1257 (UMAP)-based method) according to the dimension reduction strategy (Peres-Neto et
 1258 al., 2005; Laurens and Hinton, 2008; Andrews and Hemberg, 2018; Becht et al.,
 1259 2018).

1260 In linear dimension reduction, LDA and PCA are two widely used algorithms,
 1261 LDA distinguishes features from the aspect of the largest classification while PCA
 1262 orthogonally extracts the main components from the angle of the largest variance.
 1263 Despite the improved algorithms of JPCDA, and LDA-PLS, the dimension reduction
 1264 effect of the LDA model in single-cell transcriptome data is still not optimal (Tang et
 1265 al., 2014; Zhao et al., 2020). PCA is another linear transformation, Seurat usually
 1266 determines the amount of the PCs numbers according to the inflection point of the
 1267 standard deviation-PC diagram or the proportion test result P-value (the
 1268 ScoreJackStraw function) of the PCs. Other variants PCA based dimension reduction
 1269 methods include the pcaReduce (Žurauskienė and Yau, 2016), GLM-PCA (Townes et
 1270 al., 2019), RPCA (Gogolewski et al., 2019), tRPCA (Candès et al., 2011), scPCA
 1271 (Boileau et al., 2020), PCAone (Li et al., 2022k). GLM-PCA extends the traditional
 1272 PCA analysis to non-normal distributions, directly handles the original matrix by
 1273 introducing an exponential family likelihoods strategy to make the PCA free from
 1274 normalization restriction, and then ranks and extracts the gene implementation using
 1275 bias (Collins et al., 2002). ScPCA uses contrastive PCA and sparse PCA to remove the
 1276 technical noise and the data, respectively, which further increases the stability of the
 1277 PCA (Zou and Tibshirani, 2006; Abid et al., 2018). As most scRNA-seq datasets are
 1278 difficult to effectively represent by simple linear dimension reduction, one first
 1279 strategy for solving this is based on a rapid PCA analysis approach. PCAone proposes
 1280 a new fast randomized singular value decomposition (RSVD) strategy, which
 1281 completes the analysis of 1.3 million mice brain cells single-cell data within 35
 1282 minutes (Li et al., 2022k).

1283 Nonlinear dimensionality reduction is another solution, like non-parametric
 1284 dimension reduction methods t-SNE and UMAP, both need to set the hyperparameters
 1285 of clustering in advance; and in classification effect, the former tends to discrete the
 1286 formation of cells in the data. In the case of the reasonable use of parameters for
 1287 specification, there is no significant difference between UMAP and t-SNE, which
 1288 means after using the same method of information initialization, they can produce
 1289 approximate analytical efficiency while preserving the global structure of the data set
 1290 (Do and Canzar, 2021; Kobak and Linderman, 2021). Modified methods for t-SNE
 1291 include net-SNE, qSNE, FIt-SNE, and Joint t-SNE (Cho et al., 2018a; Linderman et
 1292 al., 2019; Wang et al., 2022b) while the improvement of the UMAP mainly comes
 1293 from the self-improvement of the method by the lmcmnes group. To better visualize
 1294 the dimension reduction results of t-SNE or UMAP, Hyunghoon Cho proposes the
 1295 den-SNE/densMAP approaches for the transcriptome variability information based on
 1296 local radius dependent optimization to iteratively optimize the function of
 1297 conventional t-SNE/UMAP; Stefan Canzar proposes the j-SNE/j-UMAP to improve
 1298 the multimodal omics data joint visualization results to reduce misleading of
 1299 visualization (Do and Canzar, 2021; Narayan et al., 2021).

1300

1301 *Clustering*

1302 In the analysis of single-cell transcriptome data, clustering is performed to divide the
 1303 cells into subgroups and we are therefore able to characterize the different cell types

1304 in multicellular organisms which helps us to accurately analyze different tissues or
 1305 developmental processes from the perspective of cell heterogeneity. The actual effect
 1306 of clustering can be affected by the pre-data processing steps like batch effect
 1307 normalization, imputation, dimension reduction, etc.

1308 After feature gene selection and dimension reduction, the vast majority of single
 1309 cells are clustered based on distance. The concept of the K-means clustering algorithm
 1310 was used for applications like SCUBA, SC3, and RaceID (Macqueen, 1967; Marco et
 1311 al., 2014; Grün et al., 2015; Kiselev et al., 2017). On parameter selection
 1312 improvement, SAIC iteratively optimizes multiple initial centers K and P-value by the
 1313 Davies-Bouldin index to obtain the optimal solution; LAK applies a parameter
 1314 selection algorithm to datasets for automatic selection of parameters (Davies and
 1315 Bouldin, 1979; Yang et al., 2017; Hua et al., 2020). In the operation of
 1316 ultra-high-dimensional data, LAK adds the Lasso penalty to make standardization and
 1317 mbkmeans achieves rapid clustering at the million-cell level using mini-batch
 1318 k-means (Hicks et al., 2021). SMSC applies a spectral clustering method to improve
 1319 the clustering performance but loses some accuracy for ultra-high-dimensional data
 1320 (Qi et al., 2021). Another broad class of widely used distance clustering methods
 1321 depends on shared the nearest neighbor graphs structure and graph cluster, among the
 1322 most widely used are Louvain or Leiden (Blondel et al., 2008; Xu and Su, 2015). The
 1323 identification of rare cells needs to be improved in combination with specific methods,
 1324 such as dropClust using the locality sensitive hashing workflow for screening the
 1325 nearest neighbor followed by Louvain cluster, it uses the exponential decay function
 1326 to retain more transcriptomic features of the rare cells (Sinha et al., 2018). Other
 1327 distance-based clustering methods use different algorithm cores: SIMLR uses a
 1328 Gaussian kernels learning model to construct kernel matrix for the potential C cell
 1329 populations in the datasets while Conos proposes a joint mutual nearest-neighbor
 1330 (mNN) graph cluster to achieve integrative analysis of multiple different single-cell
 1331 transcriptome samples (Wang et al., 2017a; Barkas et al., 2019). Density-based
 1332 clustering uses the closeness of the sample distribution for the cluster, DBSCAN is
 1333 the most classical algorithm (Fukunaga et al., 1975; Ester, 1996). For single-cell
 1334 sequencing, densityCut and FlowGrid are designed based on this principle (Ding et al.,
 1335 2016; Fang and Ho, 2021). Hierarchical clustering is a bottom-up clustering method
 1336 that repeatedly calculates cell-to-cell similarity for classification until the preset
 1337 number of clusters is completed without advance through unsupervised learning (Guo
 1338 et al., 2015). Subsequently, the RCA cluster uses a conventional hierarchical
 1339 clustering method to cluster the cells mapped to global reference panel; HGC
 1340 constructs a hierarchical tree on the SNN graph (Li et al., 2017; Zou et al., 2021). To
 1341 solve the defects that the conventional hierarchical clustering method hardly clusters a
 1342 certain group of cells and only allows the same set of signature genes for clustering,
 1343 K2Taxonomer uses the constrained k-means algorithm to expand to sample groups,
 1344 integration calculations are performed recursively based on multiple genes sets to
 1345 capture subgroups (“taxonomy-like cells”) under various resolutions (Reed and Monti,
 1346 2021). Mrtree applies hierarchical clustering’s strategy to multiple partitions of flat
 1347 cluster and constructed a multi-resolution reconciled tree to use as cell clustering

1348 (Peng et al., 2021a). Recently, Noam Kaplan proposes a KMD clustering method,
 1349 which eliminates the hyperparameter K while clustering through an average linkage
 1350 hierarchical clustering model, greatly reducing judgment errors caused by subjectivity
 1351 (Zelig and Kaplan, 2020).

1352 The deep learning cluster method is a combination of the machine learning
 1353 method and the above single-cell transcriptome clustering strategy, which can achieve
 1354 more efficient clustering results in the form of unsupervised, supervised, or
 1355 semi-supervised. These methods tend to learn a nonlinear transformation, obtaining
 1356 the best low-dimensional representation by mapping the original high-dimensional
 1357 data into a smaller latent space. Overall, this approach avoids the impact of traditional
 1358 clustering methods on the choice of pre-cluster data processing methods.
 1359 Unsupervised clustering methods include ADClust, DESC, SAUCIE, and VAE-SNE,
 1360 they usually do not require the parameters such as a preset number of clusters to
 1361 complete the analytical processing of the data set in the way of autonomous learning
 1362 (Amodio et al., 2019; Graving and Couzin, 2020; Li et al., 2020c; Zeng et al., 2022c).
 1363 Although the unsupervised clustering method avoids parameters such as manual input
 1364 cluster number and extends to ultra-high-dimensional cell clustering, sometimes using
 1365 high-quality annotated data sets or other prior knowledge auxiliary constraints for
 1366 supervised or semi-supervised clustering can achieve more accurate cell type
 1367 classification and improve clustering performance (Bai et al., 2021). Transfer learning
 1368 based ItClust, mutual supervised ZINB auto-encoder and GNN based scDSC, soft
 1369 K-means convolutional auto-encoder based ScCAEs, Cramer-World distance
 1370 max-mean penalty Gaussian mixture auto-encoder based SeGMA, time series
 1371 clustering network based STCN are all widely used supervised clustering (Hu et al.,
 1372 2020a; Ma et al., 2021b; Smieja et al., 2021; Gan et al., 2022; Hu et al., 2022a).
 1373 Furthermore, Zhang et al. have utilized hierarchical GAN to design another
 1374 widespread DL method IMDGC for single-cell transcriptome data analysis to
 1375 construct cellular embedding cluster in a generated manner (Yang et al., 2023b).

1376 For the special cases in the clustering, targeted purposes clustering methods are
 1377 designed as follows: GiniClust (Jiang et al., 2016) (updated to GiniClust 3 (Dong and
 1378 Yuan, 2020)), MicroCellClust (Gerniers et al., 2021) for rare cell subpopulations
 1379 clustering; EDClust (Wei et al., 2022), ENCORE (Song et al., 2021b) and MLG (Lu et
 1380 al., 2021) for noise reduction and batch effect removal; ClonoCluster (clonal origin
 1381 information) (Richman et al., 2023), IsoCell (alternative splicing information) (Liu et
 1382 al., 2023) clustering with additional information. Wu and Yang evaluated the cluster
 1383 methods from the perspective of the effect of feature selection on cluster, they
 1384 proved that more representative feature selection enhances the level of cell clustering,
 1385 methods based on “cluster similarity” (most distance-based clustering methods
 1386 mentioned in our review) generally have a wide range of high clustering type
 1387 performance, however, high accuracy and high running speed need targeted selection
 1388 according to the actual data set (Su et al., 2021; Yu et al., 2022). Double dipping
 1389 presents a significant issue wherein the same expression data are used both in cell
 1390 clustering and differential expression genes, resulting in an excessively high false
 1391 discovery rate (FDR) of DE genes when the cell cluster is incorrect. For example, if

1392 only one specific cell cluster is present, no gene should be considered as differential
 1393 genes. To address this problem systematically, ClusterDE adopts a cluster contrast
 1394 learning strategy for post-clustering DE testing. It demonstrates better FDR control
 1395 across different threshold ranges compared to the truncated normal (TN) test and the
 1396 Countssplit method (Song et al., 2023a).

1397

1398 *Cell typing annotation*

1399 Cell typing annotation refers to the usage of specific information to annotate cells or
 1400 cell subsets in single-cell sequencing dataset, which is the basis for subsequent
 1401 biological analysis. The most commonly used strategy is unsupervised clustering of
 1402 cells followed by annotation based on the marker genes such as scCATCH and SCSA
 1403 (Cao et al., 2020b; Shao et al., 2020a). However, it is difficult to treat complex
 1404 high-dimensional datasets (Franzén et al., 2019; Luecken and Theis, 2019; Zhang et
 1405 al., 2019c). Currently, multiple methods to automatically cell typing have been
 1406 developed and can be roughly divided into two categories, i.e., reference-dependent
 1407 and reference-free annotation methods.

1408 Reference-dependent annotation methods require users to provide pre-annotated
 1409 high-quality single-cell transcriptome datasets or prior knowledge from the
 1410 PanglaoDB database, ScType database, etc. for alignment (Ianevski et al., 2022).
 1411 According to the different principles of the method, it can be divided into
 1412 hierarchy-tree-based methods (CHETAH (de Kanter et al., 2019), Garnett (Pliner et al.,
 1413 2019), HieRFIT (Kaymaz et al., 2021), scHPL (Michielsen et al., 2021), scMRMA (Li
 1414 et al., 2022d)), similarity-based methods (SingleR (Aran et al., 2019), scmap (Kiselev
 1415 et al., 2018), deCS (Pei et al., 2023), scID (Boufea et al., 2020), scMatch (Hou et al.,
 1416 2019), Symphony (Kang et al., 2021)), signature-gene-based methods (Cellassign
 1417 (Zhang et al., 2019a), Cell-ID (Cortal et al., 2021), scMAGIC (Zhang et al., 2022g),
 1418 SciBet (Li et al., 2020b)) and other DL methods. As an early method, ACTINN is a
 1419 deep learning approach using a 3 hidden layers neural network for annotation
 1420 classification (Ma and Pellegrini, 2020). SCPred then proposes a method using
 1421 machine-learning probability-based prediction based on the unbiased feature selection
 1422 from embeddings (Alquicira-Hernandez et al., 2019). Other methods such as Seurat
 1423 project query cells in PCA space and train cell typing annotation through weighted
 1424 vote classifier; scSorter adopts a Gaussian mixture model and GraphCS uses a virtual
 1425 adversarial training (VAT) loss modified GNN to expand to multi-species, large-scale
 1426 datasets of cellular annotation (Guo and Li, 2021; Zeng et al., 2022a).

1427 Non-reference annotation approach uses a pre-trained deep learning model and
 1428 can directly perform cell classification using the query dataset as input alone.
 1429 scDeepSort uses single-cell atlas from human cell landscape (HCL) and mouse cell
 1430 atlas (MCA) database as the input for the pre-trained weighted GNN models, which is
 1431 suitable for human and mouse cell annotation with good results (Han et al., 2018b;
 1432 Han et al., 2020; Shao et al., 2021b). Similarly, Pollock is a pretrained human cancer
 1433 reference VAE model to classify the multimodal cells in the cancer environment
 1434 (Storrs et al., 2022). Although it is more convenient to use, it is difficult to achieve a
 1435 better cellular annotation effect for significantly different query datasets, and it is also

1436 difficult to expand the application due to the accuracy and the number of pre-training
 1437 reference datasets. There are also some other cell annotation tools for targeted field
 1438 research, for example, DevKidCC (Wilson et al., 2022) for human kidney cell
 1439 annotation, ikarus (Dohmen et al., 2021) for the identification of cancer and normal
 1440 cells. Overall, the performance of the non-reference annotation approach is restricted
 1441 by the coverage and accuracy of pretrained reference datasets.

1442 Currently, to improve cell annotation tools to uniformly assign cell types across
 1443 large platforms and multicell patterns is the mainstream of cellular annotation
 1444 research directions, the latest Cellar and ELeFHAnt methods have made some
 1445 attempts in this regard and achieved initial results (Thorner et al., 2021; Hasanaj et al.,
 1446 2022). Overall, similarity-based annotation methods are computationally intensive,
 1447 when applied to very large query and reference data sets, they often make a trade-off
 1448 between accuracy and speed, it is therefore generally only suitable for cell
 1449 classification in smaller datasets; for larger-scale datasets, it is recommended to use
 1450 F-test feature selection or MLP classifier (Hu et al., 2020a; Huang and Zhang, 2021;
 1451 Ma et al., 2021c). Moreover, the method of semi-supervised transfer learning such as
 1452 Itclust has good results in discovering new cell subtypes. In recent years, new
 1453 methods based on the above reference annotation method classification have been
 1454 continuously improved, and deep learning models such as VAE have also been
 1455 applied in this field.

1456

1457 *Differential expression analysis (DEGs)*

1458 Statistical tests are commonly used in the differential gene analysis of Bulk RNA-seq,
 1459 similar to the section 2.4 HVG Selection algorithm: p-values and log-fold changes are
 1460 usually used as important parameters. Statistical tests include t-test (two sample
 1461 based), Wilcoxon test, Kolmogorov Smirnov test (KS-test), and Kruskal-Wallis test
 1462 (KW-test), some of which are also widely used in the test of single-cell transcriptome
 1463 DEGs. Based on this, the corresponding detection tools are developed: limma (Ritchie
 1464 et al., 2015), edgeR (Robinson et al., 2010), and DESeq2 (Love et al., 2014). Both the
 1465 limma and edgeR algorithms are proposed by Smyth GK, the former is based on a
 1466 normal or approximate normal distribution model while the latter is based on an
 1467 overdispersed Poisson distribution model. DESeq2 is based on the NB distribution
 1468 model for hypothesis testing and uses the empirical Bayes procedure for DEGs.
 1469 Currently, limma has large errors in RNA count analysis due to specific distribution
 1470 model assumptions, although both edgeR and DESeq2 utilize the Bayes model to
 1471 normalize over-dispersion, the latter has better analysis results as promoting the
 1472 screening of CPM threshold through the average value of data set reads and outlier
 1473 detection.

1474 Single-cell transcriptome DEGs can be roughly divided into early parametric
 1475 tests on zero-value, non-parametric tests, and other methods according to time and
 1476 analytical methods. Since there are vast zero numbers in the scRNA-seq data, most of
 1477 the early methods are based on this observation to make parameter tests: such as
 1478 Monocle (Trapnell et al., 2014), SCDE (Kharchenko et al., 2014), MAST (Finak et al.,
 1479 2015), scDD (Korthauer et al., 2016), D3E (Delmans and Hemberg, 2016), TASC (Jia

1480 et al., 2017), DEsingle (Miao et al., 2018), and HIPPO (Kim et al., 2020b). The
 1481 evaluation of some methods above shows that although they generally achieve good
 1482 results in the analysis of single-cell datasets, there is no significant performance
 1483 improvement over the DEA method for bulk data (Soneson and Robinson, 2018). It is
 1484 possible that the best distribution model is not used for different datasets, and thus one
 1485 alternative solution is to consider non-parametric DEA methods.

1486 Non-parametric test or distribution-free test does not need to make prior
 1487 assumptions about the data distribution form and it is therefore applicable to the
 1488 analysis of multiple datasets, common methods are Swish (Zhu et al., 2019a), IDEAS
 1489 (Zhang et al., 2022d), ccdf (Gauthier et al., 2021), distinct (Tiberi et al., 2022). Swish
 1490 evaluates transcript level by Salmon Gibbs and then the FC value is calculated by the
 1491 Mann-Whitney Wilcoxon test. IDEAS is a pseudo F-statistic test using
 1492 Jensen-Shannon divergence (JSD) or Wasserstein distance (Was) for gene different
 1493 expression measurement, the P value is generated by PERMANOVA based distance
 1494 tester kernel based regression. Ccdf is a conditional independence test relying on the
 1495 conditional cumulative distribution function, the DEGs is predicted by a multiple
 1496 regressions model. Distinct proposes a hierarchical non-parametric permutation
 1497 method, the total distance of empirical cumulative distribution function (ECDF) is
 1498 used for DEGs identification. Alternative methods include deep learning strategies
 1499 MRFscRNAseq (Li et al., 2021a), pseudotime inference based PseudotimeDE (Song
 1500 and Li, 2021), non pre-cluster based singleCellHaystack (Vandenbon and Diez, 2020),
 1501 multiple scores based MarcoPolo (Kim et al., 2022). It is suggested that different
 1502 single-cell transcriptomic datasets should employ data-specific DE genes detection
 1503 strategies for optimal DEGs analysis, based on the scCODE workflow, the most
 1504 optimized DEGs method can be found using indicators involving CDO (DE genes
 1505 order) and AUCC (area under concordance curve) (Zou et al., 2022). In addition, the
 1506 research method will have a specific research orientation under different research
 1507 backgrounds, for example, in dose-response studies after administration DEGs
 1508 analysis, LRT linear test, and Bayesian multiple group test have better results than
 1509 other methods (Nault et al., 2022).

1510

1511 *Visualization*

1512 Single-cell transcriptome data analysis visualization refers to the visual presentation
 1513 of the above analysis results in the form of graphs, ggplot2 is the most extensive R
 1514 visualization tool and is commonly used in R to greatly enhance drawing power
 1515 (Wickham, 2009). ARL is another R package that specifically displays marker gene
 1516 Association Plots and can display its features in each cluster (Gralinska et al., 2022).
 1517 Also, there are other specific packages for marker gene visualization like Complex
 1518 Heatmap that are not described in detail here. HVG visualization is usually presented
 1519 in the form of Volcano Plot, by default, the left and right part of the graph are the
 1520 underrepresented and overrepresented genes, respectively, while the middle is the
 1521 constant gene. Enhanced Volcano is a specialized R package used for drawing a
 1522 volcano graph, and ggplot2 can also be used to achieve better results by default.
 1523 Cluster visualizations are often presented in PCA plot, t-SNE plot and UMAP plot,

1524 but it is noteworthy that the results of visualization are very deceptive, since some
 1525 small cell subpopulations may represent a large number of cells shown in the UMAP
 1526 figure. Improved methods like den-SNE/densMAP and j-SNE/j-UMAP have been
 1527 proposed to solve these problems(Macqueen, 1967; Marco et al., 2014). Furthermore,
 1528 FastProject can output a 2D display of the annotated cluster and PieParty can draw
 1529 color maps for each gene in the cluster 2D graph (DeTomaso and Yosef, 2016;
 1530 Kurtenbach et al., 2021).

1531 Meanwhile, the interactive visualization of single-cell transcriptome data is
 1532 currently a hot field, software such as Single Cell Explorer can achieve interactive
 1533 visualization to certain extent, but it is still necessary to increase the interaction
 1534 freedom to provide a more comprehensive 3D presentation of single-cell
 1535 transcriptome data (Feng et al., 2019; Cakir et al., 2020). To this end, CellexalVR uses
 1536 VR theory for interaction visualization; CellView is a Web-based tool, including the
 1537 Explore tab, Co-expression tab, Subcluster-analysis tab modules for different uses;
 1538 Cellxgene VIP is a cellxgene framework-based plugin and extends to the interactive
 1539 visualization of ST data based on combination of multiple modules (Bolisetty et al.,
 1540 2017; Legetth et al., 2021; Li et al., 2022e).

1541

1542 *Single-cell simulators*

1543 As single-cell transcriptome methods continue to expand, the pressing challenge lies
 1544 in the benchmarking, with the key issue being the requirement for stable and reliable
 1545 data, as direct sequencing of single-cell transcriptome may lack ground truth. The
 1546 realistic single-cell simulator data provided a known truth for benchmarking, allowing
 1547 training with real data while matching the characteristics of actual data. Additionally,
 1548 simulated data provide greater flexibility than real data, enabling analysts to adjust
 1549 parameters like dropout rate based on specific testing methodologies.

1550 Splatter is a two-step simulator framework that initially simulates estimated
 1551 parameters from real data and then incorporates additional parameters from users
 1552 (Zappia et al., 2017). Its six pre-designed pipeline module interfaces ensure the
 1553 repeatability of data generation. Recent updates have focused on specialization and
 1554 generalization. In the specialization domain, splaPop generates population-scale data
 1555 with genetic effects (quantitative trait loci), while dyngen simulates dynamic cellular
 1556 processes like developmental trajectories (Azodi et al., 2021; Cannoodt et al., 2021).
 1557 In the generalization field, Li's team introduced the six concepts of an ideal simulator
 1558 including authenticity, preservation of genes, capture of gene correlations, robustness,
 1559 parameter tunability, and efficiency (Sun et al., 2021; Song et al., 2023b).
 1560 Subsequently, scDesign2 is proposed to meet all six properties (Sun et al., 2021),
 1561 followed by scDesign3, addressing the gap in single-cell omics statistical simulation
 1562 (Song et al., 2023b). The increased accuracy and transparency of the simulator
 1563 enhance benchmarking between different single-cell data processing methods, guiding
 1564 the selection of the most appropriate approach for specific data and licensing needs.

1565

1566 *Extended downstream analysis*

1567 *Pseudotime*

1568 In order to more truly restore the real process in the organism, integration of multiple
 1569 transcriptome data using pseudo-timing analysis is needed to reconstruct cellular
 1570 developmental trajectories by inferring cell information at different time points,
 1571 including state, distribution, number and gene expression (Bar-Joseph et al., 2012;
 1572 Bendall et al., 2014; Ding et al., 2022). This dynamic analysis of transcriptome
 1573 features is known as Pseudotime analysis (Figure 4E). Based on whether it depends
 1574 on gene expression, pseudotime analysis methods can be divided into gene (exons)
 1575 expression-based method and RNA-velocity-based method.

1576 Pseudotime analysis based on gene expression level was first proposed, it usually
 1577 clustering methods such as dimensional reduction is used to construct multi-branching
 1578 graphs model in a low-dimensional space to mimic the developmental trajectory of
 1579 cells: minimal spanning tree (MST) based method monocle (Trapnell et al., 2014),
 1580 monocle 2 (Qiu et al., 2017), TSCAN (Ji and Ji, 2016); PAGA based method PAGA
 1581 (Wolf et al., 2019), monocle 3 (Cao et al., 2019); other graph architectures method
 1582 Wishbone (Setty et al., 2016), p-Creode (Herring et al., 2018) are all for this purpose.
 1583 MST is a model that connects all points in a 2-dimensional plane and has the lowest
 1584 total connection weight, it was used first to solve the traveling salesman problem, Qiu
 1585 et al. applied MST model constructed with Boruvka's algorithm to analyze cellular
 1586 hierarchy in 2011 (Qiu et al., 2011). Monocle maps cells into a high-dimensional
 1587 Euclidean space and reduces dimension using ICA, Monocle 2 updates the monocle
 1588 and uses the reversed graph embedding (RGE) strategy to construct cell path, cells are
 1589 subsequently distributed to the spanning tree constructed using centroids. PAGA
 1590 (partition-based graph abstraction) preserves the global topology structure of the
 1591 dataset, by statistical connectivity measures of the neighborhood graph weights (KNN
 1592 graph by default), PAGA graphs at multiple resolutions are produced to conduct
 1593 pseudotime analysis based on an expanded diffusion pseudotime (DPT) method.
 1594 Monocle 3 combines the advantages of monocle 2 and PAGA to form multiple PAGA
 1595 graphs on the UMAP space, then uses the SimplePPT algorithm to learn the principal
 1596 graph and then constrained by other PAGA graphs, the final derived cell
 1597 developmental trajectories can be adapted to large datasets with compositional
 1598 complexity. Overall, PAGA and monocle 3 comprehensively consider the
 1599 computational speed, accuracy, and robustness, and are currently the best methods for
 1600 the pseudotiming analysis of the single-cell transcriptome. In addition to the graph
 1601 method, other gene expression based methods include CSHMMs which use HMM
 1602 model to calculate the distance between each cell to root cell and then complete cell
 1603 trajectory assignments iteratively; SCUBA which uses a bifurcation analysis model;
 1604 SLICE which proposes a scEntropy directed model as highly differentiated cells have
 1605 minimized scEntropy (Marco et al., 2014; Guo et al., 2017; Lin and Bar-Joseph,
 1606 2019).

1607 The RNA-velocity-based method relies on the content of RNA velocity which is
 1608 proposed first by Peter V. Kharchenko in 2018, they think that the ratio of
 1609 unspliced/spliced mRNA can be used in infer transcriptional dynamics as cells with a
 1610 higher proportion of uncleaved mRNA are younger (as a later cell differentiation state)
 1611 (La Manno et al., 2018). At the same time, they also propose a dedicated analysis

software, velocyto (available through the R package of the velocyto.R) as a steady-state model to quantify RNA velocity for developmental trajectory analysis. scVelo is another analysis tool designed specifically for RNA velocity, it uses the likelihood-based dynamical model to solve the cell trajectory inference with steady-state mRNA levels and situation violates the central assumption of the common splicing rate (Bergen et al., 2020). But there is still room for methodological improvement in velocity projection methods: constant degradation and nuclear export assumptions still need to be proved. This also provides a direction for the subsequent RNA velocity based method (Bergen et al., 2021). Methods concerning deep learning have been widely used in the modeling prediction of RNA velocity to further enhance processing power for large-complex datasets, like the Bayesian hierarchical model BRIE2 (Huang and Sanguinetti, 2021); velocity auto-encoder model based VeloAE (Qiao and Huang, 2021); variational auto-encoder model DeepCycle (Riba et al., 2022).

1626

1627 *Cell-cell interaction*

1628 Cell-Cell Interaction (CCI) is an important feature for maintaining the normal
 1629 physiological function in multicellular organisms, which determines the fate of cells
 1630 exploring the mechanism of disease occurrence, exploring the genetic variation
 1631 process and other regulatory processes (Singer, 1992; Shao et al., 2020b). Cell
 1632 interaction network intuitively embodies the interaction relationship between cells
 1633 (Figure 4F).

1634 Direct CCI based on the neighborhood structure refers to the extraction and
 1635 analysis of the CCI with a possible direct contact, using the physical distance between
 1636 the cells. ProximID method completes the physical cellular network construction on
 1637 eligible cells with a predetermined interaction distance (Euclidean distance) (Boisset
 1638 et al., 2018). Neighbor-seq identifies the cell types using a random forest classifier,
 1639 CCI network is constructed by the igraph method using enrichment scores calculation
 1640 score (Csardi and Nepusz, 2006; Ghaddar and De, 2022). Due to the great limitations
 1641 of this analysis method, it is not currently used alone. This KNN connected graph is
 1642 commonly used as one of the inputs to the CCI for deep learning, and its physical
 1643 distance has also become an important hypothesis in single-cell CCI studies (two
 1644 adjacent cells physically in direct contact are more likely to have some form of
 1645 interaction than two random cells) for global CCI analysis.

1646 The complete process of CCI relationships for indirect contact should include
 1647 ligands, receptors, signaling proteins, transcriptional factors (TFs), and target genes.
 1648 Common indirect CCI methods mainly use a priori ligand-receptor pairs databases
 1649 (like the CellTalkDB database which integrates information from validated 3,398
 1650 human LR pairs and 2,033 mouse LR pairs (Shao et al., 2021a)), a cell-cell
 1651 connection matrix is made in which each value represents the co-expression level of
 1652 LR pairs. Then a cell connection graph is constructed for CCI analysis, the main
 1653 analytical method packages include single-cell CCI inference method SoptSC (Wang
 1654 et al., 2019d), Scriabi (Wilk et al., 2023); LR pairs based cluster CCI method NATMI
 1655 (Hou et al., 2020), SingleCellSignalR (Cabello-Aguilar et al., 2020), scCrossTalk

(Shao et al., 2024), CellPhoneDB (Efremova et al., 2020), Nichenet (Browaeys et al., 2020), CellChat (Jin et al., 2021), CellCall (Zhang et al., 2021d), ICELLNET (Noël et al., 2021), scMLnet (Cheng et al., 2021), CytoTalk (Hu et al., 2021b), Tensor-cell2cell (Armingol et al., 2022), LRLoop (Xin et al., 2022); other information based clusters CCI method InterCellDB (Jin et al., 2022b), EBOCOST (Zheng et al., 2022b). The LR pair-based approach constructs the database using the literature database curated or previous self-validated LR pairs: NATMI uses the connectomeDB2020 database by default (1751 of 2293 LR pairs were from the validated draft map by author in 2015) to construct weighted directed multi-edge networks (Ramilowski et al., 2015). CellPhoneDB proposes a certain SQLite database to retain specific subunit architecture of LR pairs, mean expression threshold is used to determine the interacting cells, and a geometric sketching subsample framework is used for enhanced power to large datasets and excluded noise. As similarly, ICELLNET takes use of the multi-subunit structure of ligands with receptors for heteromeric complexes. NicheNet uses model-based parameter optimization on an LR prior model to optimize CCI intensity by adding intracellular signaling information (target gene). It overcomes the problem that the above methods directly use the receptor gene expression level to represent the amount of receptor protein in the cells and combine the downstream signaling pathway with GRN to improve CCI analysis. Therefore, in the analysis of the single-cell transcriptome CCI, the CellPhoneDB and NicheNet are usually used together to achieve the best analysis results (Dimitrov et al., 2022).

The latest methods of single-cell CCI adopt the strategy of DL and improve the application performance to some extent. DeepLinc uses a VGAE model to reconstruct full range intercellular CCI network (Li and Yang, 2022). TraSig is a continuous-state Hidden Markov Model that uses pseudotime ordering to calculate dynamic interaction scores for CCI inference (Li et al., 2022a). In addition, as now spatially resolved transcriptomics (ST) provides the gene information with crucial spatial information, the inference of spatial cell-cell communications remains a great challenge. SpaOTsc can reconstruct the spatial properties of scRNA-seq data and build the CCI network relying on a structured optimal transport method (Cang and Nie, 2020). Giotto uses a cell-cell proximity graph to infer the signaling pathways (Dries et al., 2021b). However, both SpaOTsc and Giotto hardly resolve the spot-based ST data. Recently, Fan's lab presents SpaTalk which uses a knowledge graph and graph network to construct a ligand-receptor-target network between spatially adjacent cells for both single-cell and spot-based ST data (Shao et al., 2022a).

1691

1692 *Pathway enrichment analysis*

1693 Gene pathway enrichment analysis refers to using the gene of interest as a foreground
 1694 gene and known specific database associations to establish gene-biological process
 1695 links which are used to explain the physiological functions of differentially expressed
 1696 genes, upstream and downstream pathways, etc (Creixell et al., 2015). Gene Ontology
 1697 (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) are the first batch
 1698 proposed databases used as an enrichment analysis (Ogata et al., 1999; Ashburner et
 1699 al., 2000). Gene Set Enrichment Analysis (GSEA) is another widely used method

1700 calculating the enrichment score to determine whether gene set S will occur in both
 1701 sides of ranker DEGs List L as well as the significance test value (Subramanian et al.,
 1702 2005). All pathways in Ingenuity Pathway Analysis (IPA) software have been
 1703 experimentally verified and it can predict the changing trend of the entire pathway
 1704 when activated compared with other analysis. Other common database pathway
 1705 enrichment methods also include Over-Representation Analysis (ORA) (Khatri et al.,
 1706 2012), Network Topology-based Analysis (NTA) (Wang et al., 2013), Reactome gene
 1707 sets (Fabregat et al., 2018), CORUM complexes (Ruepp et al., 2010). The difference
 1708 in the source pathways and enrichment means of the reference datasets will directly
 1709 affect the pathway enrichment results. Web-based online analysis tools with different
 1710 integrated databases have been proposed to easier analysis (Zhang et al., 2005; Wang
 1711 et al., 2017b). Metascape involves the transcriptome databases (KEGG, GO, CORUM,
 1712 TRRUST, etc.) and protein-protein interactions databases (STRING, BioGrid,
 1713 OmniPath, etc.), there are a total of 25 databases that can be used for genetic and
 1714 proteomic enrichment analysis in 8 species including human and mouse (Zhou et al.,
 1715 2019b).

1716 In conclusion, although we have enumerated the most common parts of
 1717 single-cell transcriptome downstream analysis (Table S3 & Table S4), there are still
 1718 many methods not covered including gene regulatory network analysis, immune
 1719 analysis, cell cycle assignment, gene variants exploration, alternative splicing analysis,
 1720 etc. Overall, single-cell transcriptome analysis methods are diverse and still evolving,
 1721 the starting point and the ultimate goal of all the analysis methods is to use the
 1722 accurate mining of the biological information from the single-cell transcriptome
 1723 sequencing data for biological explanation.

1724

1725 *Applications of scRNA-seq*

1726 ScRNA-seq has become a potent instrument, empowering scientists to delve into the
 1727 intricate realm of individual cells and unveil their distinct molecular characteristics.
 1728 Leveraging scRNA-seq, investigators now have the capability to delve into cellular
 1729 diversity, study developmental biology, scrutinize disease progression, and advance
 1730 drug development with unparalleled precision. This methodology has unlocked fresh
 1731 possibilities for revealing novel biomarkers, pinpointing therapeutic targets, and
 1732 forging pathways toward personalized medicine. In this segment, we illuminate and
 1733 deliberate on several noteworthy applications of scRNA-seq in the realms of
 1734 biomedical and clinical research.

1735

1736 *Application of scRNA-seq in embryonic, tissue, and organ development research*

1737 *Embryonic development research*

1738 ScRNA-seq proves to be pivotal in embryonic development, particularly in the
 1739 identification and categorization of distinct cell types and lineages.

1740 In a research by Cao et al., 38 primary cell types and 655 subtypes were
 1741 identified by scRNA-seq analysis of 2,072,011 single cells from mouse embryos (Cao
 1742 et al., 2019). This comprehensive information sheds light on the developmental
 1743 trajectories of different cell types during organogenesis in mammals, culminating

in the creation of a developmental-specific trajectory map for skeletal muscle cells. This research significantly contributes to advancing our knowledge in the realm of mammalian developmental biology. Moreover, researchers leverage scRNA-seq to reconstruct developmental trajectories, unraveling the regulatory networks that govern cell fate decisions through the analysis of gene expression at different developmental stages (Wu et al., 2024). Scialdone et al.'s study, which analysed 1,205 mouse cells from the early gastrula formation stage, used single-cell transcriptome analysis, created gene expression profiles for healthy mammals during the early developmental stage, and investigated the function of the important hematopoietic transcription factor Tal1, serves as an instructive example (Scialdone et al., 2016). Nestorowa et al. harnessed the potential of scRNA-seq to profile over 1,600 single hematopoietic stem and progenitor cells, unveiling the trajectories of hematopoietic stem cells and shedding light on the molecular events orchestrating blood cell differentiation (Nestorowa et al., 2016). Additionally, scRNA-seq also plays an essential roles in identifying pivotal regulatory factors and signaling pathways dictating cell fate decisions during embryogenesis. In a study by Lescroart et al., Mesp1 emerged as a crucial transcription factor involved in the specification of cardiac progenitor cells during mouse heart development, revealed through scRNA-seq (Lescroart et al., 2018). By integrating single-cell transcriptomic analyses of human mesoderm cells derived from embryonic stem cells and embryos, Wen et al. identified and defined the molecular characteristics of human hematopoietic mesoderm cells biased towards hematopoietic lineages (Wen et al., 2024). Cui et al. (Cui et al., 2019) used scRNA-seq to explore the gene expression landscapes of almost 4,000 cardiac cells from human embryos. They identified four main cell types and uncovered important signalling pathways that may be essential for the maturation and differentiation of several cell types. This wealth of information lays the foundation for a deeper understanding of the mechanisms governing *in vivo* human cardiac development.

1771

1772 *Tissue and organ development research*

1773 Apart from studying embryonic cells, scRNA-seq is a useful method to describe the
 1774 evolution of certain cell populations in organs or tissues, providing information about
 1775 important transitions in the development of animal tissues and organs (Mu et al., 2019;
 1776 Paik et al., 2020). Subsequently, we will delve into the applications of scRNA-seq in
 1777 tissue and organ research, emphasizing its significance in unveiling cellular diversity,
 1778 identifying rare cell populations, and comprehending disease mechanisms.

1779

In 2016, the Human Cell Atlas (HCA) project was initiated as a large-scale international collaboration with the goal of mapping and characterizing all distinct cell types in the human body. The primary objectives include understanding the spatial organization and functional relationships of these cell types, advancing our knowledge of human biology, and ultimately enhancing the diagnosis and treatment of diseases. Since its inception, numerous scRNA-seq analyses have contributed to depicting the landscapes of human cells. Using microwell-seq, Han et al. (Han et al., 2020) profiled 702,968 single cells from seven different types of cell culture and 60 human tissue types, revealing cell heterogeneity in a variety of human tissues that had

not been known before. For example, their examination of human kidney tissues in the fetal and adult stages identified a new intercalated cell-principal cell type in the adult kidney and previously unidentified kinds of S-shaped body cells in the fetal kidney. This construction of a human cell landscape at the single-cell level serves as a valuable resource for advancing our understanding of human biology. Similarly, studies by Jones et al. (Jones et al., 2022b), Eraslan et al. (Eraslan et al., 2022), Domíguez Conde et al. (Domíguez Conde et al., 2022), as well as Suo et al. (Suo et al., 2022), reported pan-tissue single-cell transcriptome atlases encompassing more than 500 different cell types and over a million cells from 68 different donors in more than 30 human tissues. These researches identified unusual cell types, tissue-agnostic traits, tissue-specific cell states, and even disease-associated cell types through cross-tissue comparisons of cell types and their transcriptional properties. These pan-tissue investigations mark a significant milestone in constructing a comprehensive human single-cell atlas.

In addition to constructing organ atlases, scRNA-seq has proven instrumental in unveiling cell type-specific gene expression changes associated with various diseases, offering crucial insights into disease mechanisms. Wilson et al. (Wilson et al., 2019) performed unbiased snRNA-seq on cryopreserved human diabetic kidney samples. They found that key inflammatory markers TNFRSF21 and ILR1 were significantly increased in infiltrating immune cells, potentially serving as biomarkers for disease progression or targets for early intervention in diabetic kidney disease. In another study, Koenig et al. (Koenig et al., 2022) used a combination of snRNA-seq and scRNA-seq to examine cardiac tissue from 27 healthy donors and 18 people with dilated cardiomyopathy. They deciphered transcriptional programmes distinct to each major cardiac cell type, found gene expression profiles linked with the condition, and illuminated the molecular mechanisms behind dilated cardiomyopathy.

In summary, scRNA-seq has played a pivotal role in constructing comprehensive atlases of human tissues and organs. It has unveiled intricate gene expression networks and developmental trajectories, offering insights into potential factors and targets implicated in the pathogenesis of individual organs or tissues. This wealth of information has the potential to unravel disease mechanisms and provides a solid foundation for advancing disease treatment strategies.

1820

1821 *Application of scRNA-seq in tumor biology research*

ScRNA-seq, by enabling the comprehensive analysis of the entire transcriptome at a single-cell resolution, has transformed our comprehension of tumor biology. It has been instrumental in unveiling the heterogeneity within tumors, identifying distinct subclones, characterizing interactions between tumor and immune cells, revealing signaling pathways associated with tumors, and predicting responses and resistance to drugs. This technology allows for the creation of detailed cell maps, the discovery of novel biomarkers, and the identification of therapeutic targets. As scRNA-seq technology continues to advance, it holds significant promise for enhancing patient outcomes and expediting the development of personalized treatments.

1831

1832 *Tumor heterogeneity*

1833 Tumors exhibit a diverse array of cell types, each characterized by unique gene
 1834 expression profiles and functional attributes. Although traditional bulk RNA
 1835 sequencing is very informative, it provides an averaged gene expression profile of the
 1836 entire tumor, masking the internal cellular heterogeneity. scRNA-seq technology is
 1837 capable of identifying and characterizing distinct cell populations within tumors,
 1838 offering a nuanced perspective of the tumor ecosystem.

1839 For instance, Hu et al. employed scRNA-seq to identify six subtypes of fallopian
 1840 tube epithelium cells in healthy human fallopian tube tissues, as well as serous
 1841 ovarian cancer (SOC) subgroups linked to patient prognosis, and reveal intra-tumoral
 1842 heterogeneity in SOC (Hu et al., 2020b). Another study by Liang et al. analyzed
 1843 scRNA-seq data from eight high-grade SOC cases and identified 20 tissue-specific
 1844 cell clusters (Liang et al., 2021a). The study developed a two-gene signature
 1845 prognostic stratification approach (CXCL13 and IL26) to identify prospective targets
 1846 for immunotherapy and precisely assess prognostic risk by taking use of the
 1847 heterogeneity of ovarian cancer immune cells. Tumor heterogeneity poses a
 1848 significant challenge to the precise diagnosis and treatment of gastric adenocarcinoma
 1849 (GA). Zhang et al. analysed 27,677 cells from nine GA samples and three non-tumor
 1850 samples using unbiased scRNA-seq technology. This investigation revealed
 1851 differentiation and cellular heterogeneity both within and across GA patients,
 1852 providing insight into the molecular makeup of an uncommon chief cell-predominant
 1853 GA type (GA-FG-CCP). The authors proposed a biomarker panel for distinguishing
 1854 between benign and malignant epithelium based on their findings (Zhang et al.,
 1855 2021b). Zhong et al. utilized scRNA-seq to explore cellular heterogeneity and regular
 1856 networks in 9 patients with multiple myeloma. Through analysis, they discovered
 1857 unique molecules, networks, and crosstalk pairs in different stages of the disease,
 1858 offering valuable insights into its prognosis and treatment (Zhong et al., 2022).
 1859 Therefore, by unraveling tumor heterogeneity, scRNA-seq aids researchers in
 1860 comprehending the various cell types present, their interactions, and their
 1861 contributions to tumor development, progression, and responses to treatment.

1862

1863 *Tumor microenvironment*

1864 The tumor microenvironment (TME) constitutes an intricate ecosystem comprising
 1865 diverse cell types, including cancer cells, immune cells, stromal cells, and vascular
 1866 cells. It holds a pivotal role in tumor development, progression, and responses to
 1867 therapy. ScRNA-seq can unravel the complexity of TME by identifying and
 1868 classifying distinct cell populations based on their gene expression profiles.

1869 Through the profiling of individual cell transcriptomes, scRNA-seq allows the
 1870 identification of rare cell populations within the TME, such as tumor-infiltrating
 1871 lymphocytes (TILs), cancer-associated fibroblasts (CAFs), and myeloid-derived
 1872 suppressor cells (MDSCs). An exemplary study conducted by a research group from
 1873 China utilized scRNA-seq to analyze the transcriptomes of 47,304 cells from nine
 1874 patients with gastric cancer. The study unveiled multiple immune cell subsets,
 1875 including regulatory T cells, CD4⁺ T cells, CD8⁺ T cells, natural killer cells, and

1876 innate lymphocyte cells (Li et al., 2022j). Notably, the study found an enrichment of
1877 regulatory T cells in gastric tumor tissues, marked by increased expression of immune
1878 suppression-related genes like DUSP4, IL2RA, TNFRSF4, LAYN, and LGALS1,
1879 indicating an immunosuppressive microenvironment in gastric tumors.

1880 Understanding TME cellular heterogeneity and gene expression may lead to the
1881 creation of novel targeted cancer therapeutics as well as cutting-edge early diagnostics.
1882 Significant heterogeneity in the infiltrating T cell population was found in a study by
1883 Savas et al., which involved the analysis of 6,311 intratumoral T cells extracted from
1884 123 breast cancer patients using scRNA-seq. According to the study, individuals with
1885 breast cancer who had a high TIL count had CD8⁺ T cells that had characteristics of
1886 tissue-resident memory T (TRM) cell development. Moreover, these CD8⁺ TRM cells
1887 exhibited significant quantities of cytotoxic effector proteins (PRF1 and GZMB) and
1888 immunological checkpoint molecules (PDCD1 and CTLA4) (Savas et al., 2018).
1889 Moreover, in early-stage triple-negative breast cancer, the gene profiles found inside
1890 the CD8⁺ TRM cluster were significantly linked to favorable patient survival. This
1891 highlights the ability of scRNA-seq to detect small subpopulations of TILs that are
1892 associated with immune surveillance or immunosuppression. These various immune
1893 cell types may be used as therapeutic targets or prognostic variables for breast cancer.
1894

1895 *Therapeutic selection and monitoring*

1896 The development of tailored treatment plans is a significant use of scRNA-seq in
1897 cancer research. The population of cells that make up tumours is diverse and includes
1898 endothelial, stromal, immunological, and malignant cells. These cell types can all play
1899 a role in treatment resistance, metastasis, and tumour formation. By analyzing the
1900 gene expression profiles of individual cells within a tumor, scRNA-seq can help
1901 identify specific cell populations that play key roles in tumor progression and therapy
1902 resistance.

1903 In a study conducted by Tirosh et al., scRNA-seq was employed to analyze the
1904 heterogeneity of melanoma tumors and pinpoint distinct cell states linked to therapy
1905 resistance (Tirosh et al., 2016). The investigation revealed a specific subpopulation of
1906 tumor cells characterized by elevated AXL gene expression, which was associated
1907 with resistance to targeted therapies. This discovery paved the way for the
1908 development of combination therapies targeting both the AXL pathway and the
1909 targeted therapy pathway, resulting in improved treatment responses. Not every
1910 patient responds to immune checkpoint inhibitors, such as anti-PD-1 and
1911 anti-CTLA-4 antibodies, which have revolutionised cancer treatment by increasing
1912 the immune system's capacity to identify and fight cancer cells. To unravel the
1913 underlying resistance mechanisms and enhance treatment outcomes, researchers at the
1914 Broad Institute of MIT and Harvard conducted a study using scRNA-seq to analyze
1915 the gene expression profiles of individual cells within tumor samples from 33
1916 melanoma patients (Jerby-Arnon et al., 2018). The scRNA-seq analysis unveiled a
1917 distinct subset of cancer cells known as the T cell exclusion program (TEX). These
1918 TEX cells actively suppressed the recruitment and activation of T cells in the tumor
1919 microenvironment, forming an immunosuppressive barrier that shielded cancer cells

1920 from immune attack. The TEX program was associated with resistance to immune
 1921 checkpoint blockade therapies and poor response to anti-PD-1 treatment and may
 1922 serve as a potential therapeutic target to overcome immune resistance.

1923

1924 *Application of scRNA-seq in immune system research*

1925 The immune system, which is made up of immune molecules, immune cells, and
 1926 immunological organs, is a crucial component of the body's internal environment. Its
 1927 job is to identify and eliminate antigenic foreign substances from the body (See et al.,
 1928 2018; Akar-Ghibril, 2022). The immune system may produce autoantigenic reactions
 1929 in the course of combating infections, which can result in immunological disorders
 1930 and harm tissues or organs (Li et al., 2022l; Suo et al., 2022). The complexity and
 1931 diversity of immune illness mechanisms make the timely identification of disease
 1932 triggers essential for the treatment of immunological diseases (Zhao et al., 2021b). As
 1933 a powerful technology, scRNA-seq can discover new cell subpopulations, reveal the
 1934 developmental lineage of immune cells, and identify the regulatory programs of
 1935 immune responses in immune diseases, thereby further elucidating the pathogenesis
 1936 of immune diseases at the single-cell level and exploring new therapeutic strategies to
 1937 benefit more patients.

1938

1939 *Research on immune cell heterogeneity*

1940 ScRNA-seq can characterize individual cells within tissues and organs and identify
 1941 rare and previously unknown cell populations. In immune system diseases, this
 1942 technique has unveiled distinct immune cell subsets and their functional states.

1943 He et al. analyzed 26,456 immune cells from old zebrafish brains by scRNA-seq
 1944 and revealed the crucial role of microglia and T cells in the neurodegenerative process
 1945 in aging (He et al., 2023). A study from the USA analyzed the transcriptomes of
 1946 approximately 276,000 single peripheral blood mononuclear cells (PBMCs) from 33
 1947 children with systemic lupus erythematosus (cSLE) and 11 matched healthy controls
 1948 using scRNA-seq (Nehar-Belaid et al., 2020). This investigation identified two novel
 1949 cell subpopulations (ISG^{hi} T-SC4 and $CD8^+$ T cells expressing high levels of
 1950 cytotoxic proteins) and revealed SLE-restricted activated NK cells and ISG^{hi} NK-SC
 1951 associated with disease severity. This comprehensive profiling of SLE heterogeneity
 1952 at the single-cell level contributed to a deeper understanding of the cellular
 1953 composition and functional diversity within the immune system, shedding the
 1954 underlying mechanisms driving disease progression. Zheng et al. (Zheng et al., 2022a)
 1955 obtained the single-cell landscape associated with lupus pathogenesis by scRNA-seq.
 1956 The study elucidated the heterogeneous characteristics present in cutaneous lesions
 1957 between discoid lupus erythematosus (DLE) and SLE, contributing to a better identify
 1958 potential avenues for therapeutic intervention.

1959

1960 *Research on the mechanism of immune disease*

1961 ScRNA-seq technology empowers researchers to dissect the intricate cellular
 1962 composition of immune diseases, pinpoint dysregulated pathways, and uncover novel
 1963 cell types or subtypes that may contribute to disease pathogenesis. The utilization of

1964 scRNA-seq promises substantial progress in the diagnosis, treatment, and
 1965 management of immune diseases in the future. Using 10x Genomics, Gaydosik et al.
 1966 focused on 3,729 CD3+ lymphocytes from skin biopsies of 10 healthy donors and 27
 1967 patients with active systemic sclerosis (SSc) (Gaydosik et al., 2021). This study
 1968 revealed different tissue-resident and circulating T cell subpopulations in both healthy
 1969 and SSc skin and identified the cytokines that contribute to inflammatory immune
 1970 disorders. The findings advance our understanding of the immunological mechanisms
 1971 underlying disease processes and hold potential for the development of novel, tailored
 1972 therapy approaches in SSc. Xu et al. (Xu et al., 2022) analyzed the pathogenic
 1973 mechanism of vitiligo through single-cell transcriptome technology, revealed the
 1974 relationship between skin fibroblasts and vitiligo, and further clarified the location
 1975 preference of vitiligo onset, which has important guiding significance for the
 1976 development of new therapeutic strategies for the treatment of vitiligo.

1977

Application of scRNA-seq in infectious diseases research

1978 scRNA-seq has revolutionized the field of infectious disease research by enabling the
 1979 study of host-pathogen interactions, characterizing the host immune response, and
 1980 investigating the impact of infectious diseases on host tissues. With the use of this
 1981 technology, the complexity of infectious diseases has been untangled at the single-cell
 1982 level, revealing hitherto unexplored insights into the variety and functional states of
 1983 cellular responses during infection.

1984 A variety of diseases, including interstitial pneumonia with consolidation,
 1985 granulomatous lesions with non-necrotic or caseous necrotic centres, and cavitary
 1986 liquefied lesions, are indicative of *Mycobacterium TB* infection, which results in
 1987 pulmonary tuberculosis (Hunter et al., 2014). Six tuberculosis patients' lung tissues
 1988 were subjected to scRNA-seq by Wang et al. with the goal of investigating the
 1989 heterogeneity and intercellular interaction in regions with 18F-FDG avidity and
 1990 nearby uninvolved tissues (Wang et al., 2023b). The scRNA-seq analysis identified a
 1991 total of 29 distinct cell subsets, encompassing both immune and parenchymal cells,
 1992 each characterized by specific marker genes. The detailed characterization of these
 1993 cell types and their associated marker genes offers a comprehensive understanding of
 1994 the distinct immune and non-immune populations present in tuberculosis-infected
 1995 lungs. This information is crucial for deciphering the complex interactions between
 1996 these cells during tuberculosis infection and may aid in identifying potential new
 1997 therapeutic targets. Zhao et al. employed scRNA-seq to comprehensively analyze the
 1998 gene expression profiles of immune cells in draining lymph nodes responding to *Y.*
 1999 *pestis* infection, which may contribute to understanding of the plague pathogenesis
 2000 during the early stage of infection (Zhao et al., 2023). Chua and colleagues utilized
 2001 scRNA-seq on nasopharyngeal and bronchial samples from 19 patients with
 2002 COVID-19 to identify molecular correlates of disease severity (Chua et al., 2020). By
 2003 better understanding the underlying molecular pathways behind infectious diseases,
 2004 researchers may be able to design diagnostic tools, treatments, and preventive
 2005 measures that are more successful.

2006

2008 *Application of scRNA-seq in drug discovery and development research*

2009 The drug discovery process is often hindered by inefficiencies due to a limited
2010 understanding of human biology, including cellular heterogeneity, disease
2011 mechanisms, drug responses, and therapeutic targets. Since its inception in 2009,
2012 scRNA-seq technology has come a long way and offers a promising solution for drug
2013 development. This method can be integrated at different stages of the drug discovery
2014 and development pipeline, as it can capture individual cell whole-transcriptome
2015 profiles. During the initial phases, scRNA-seq can assist in discovering novel cellular
2016 and molecular targets. By deepening our understanding of diseases through subtyping
2017 based on altered cell compositions and states, scRNA-seq contributes to a more
2018 nuanced comprehension of pathological mechanisms. Furthermore, this technique
2019 provides insight into the actions of compounds that are particular to cell types,
2020 off-target effects, and heterogeneous responses, all of which help in the process of
2021 choosing potential new drugs. During clinical development, scRNA-seq plays a
2022 pivotal role in identifying biomarkers for patient stratification. It helps unravel drug
2023 mechanisms of action or resistance and allows for the monitoring of drug responses
2024 and disease progression. By providing insights into the cellular and molecular
2025 landscape, scRNA-seq serves as a valuable tool in optimizing the drug discovery
2026 process, ultimately facilitating the development of more effective and targeted
2027 therapeutics.

2028

2029 *Target identification*

2030 ScRNA-seq has brought about a paradigm shift in the initial stages of drug discovery,
2031 particularly in the identification of therapeutic targets crucial for disease pathogenesis.
2032 This technology facilitates target identification by unveiling dysregulated cell types
2033 and states in disease conditions. Profiling the transcriptomes of individual cells allows
2034 the identification of specific genes and pathways that exhibit differential expression in
2035 disease-associated cell populations compared to healthy counterparts. This
2036 information becomes instrumental in guiding the selection of potential therapeutic
2037 targets and prioritizing candidate molecules for further exploration.

2038 For instance, Abdelfattah et al. used scRNA-seq to analyse over 200,000 human
2039 glioma, immunological, and stromal cells at the single-cell level in glioblastoma.
2040 They discovered S100A4 to be a novel target for immunotherapy in glioblastoma
2041 using this method. Interestingly, eliminating this target in non-cancerous cells showed
2042 an amazing capacity to rewire the immune system, resulting in a notable increase in
2043 survival (Abdelfattah et al., 2022). In the context of chronic pancreatitis, single-cell
2044 sequencing of pancreatic immune cells and T cell receptors has shed light on potential
2045 therapeutic targets. The identification of the CCR6-CCL20 signaling pathway in
2046 genetic chronic pancreatitis opens avenues for targeted interventions in humans (Lee
2047 et al., 2022). In order to examine the relationship between tumours and surrounding
2048 immune cells, a study from the University of Texas MD Anderson Cancer Centre,
2049 USA, used scRNA-seq on 186,916 cells from 5 early-stage lung adenocarcinomas and
2050 14 multi-region normal lung tissues (Sinjab et al., 2021). The results of this study
2051 indicate that CD24 expression in tumour epithelium is dramatically elevated and is

2052 connected with pro-tumor immune phenotypes and decreased survival. These findings
2053 imply that CD24 could be a promising target for the treatment of early-stage lung
2054 adenocarcinome.

2055

2056 *Drug screening and optimization*

2057 ScRNA-seq has played a pivotal role in enhancing the efficiency and precision of
2058 drug screening and optimization. Traditional screening methods often rely on cell
2059 populations that may not fully capture the heterogeneity present in the target tissue or
2060 organ. Leveraging scRNA-seq, researchers can identify and isolate specific cell types
2061 or subpopulations of interest, allowing for a more nuanced assessment of their
2062 response to various drug candidates.

2063 Cao et al.'s high-throughput single-cell RNA and VDJ sequencing of
2064 antigen-enriched B cells from 60 recovering patients serves as an instructive example.
2065 Using this method, they quickly isolated 14 strong neutralising antibodies against
2066 SARS-CoV-2 from a large collection of 8,558 IgG1⁺ clonotypes that bind to antigen
2067 (Cao et al., 2020a). Among the antibodies against SARS-CoV-2, BD-368-2 was found
2068 to have the strongest neutralising impact. Additionally, its therapeutic and preventive
2069 activity was verified in hACE2-transgenic mice infected with SARS-CoV-2. This
2070 work shows how human neutralising antibodies can be effectively discovered by
2071 high-throughput single-cell sequencing, especially during pandemics of infectious
2072 diseases.

2073

2074 *Drug mechanisms of action*

2075 ScRNA-seq offers a valuable tool to gain insights into the cellular and molecular
2076 changes induced by drugs, enabling a comprehensive characterization of their
2077 mechanisms of action. An example of this application is seen in the work of Taukulis
2078 et al., who employed scRNA-seq to investigate acute cisplatin-induced ototoxicity in
2079 a mouse model. By comparing the transcriptomes of cisplatin-treated adult stria
2080 vascularis with unperturbed adult stria vascularis, the researchers identified cell
2081 type-specific regulatory networks. Their findings highlighted that marginal and
2082 intermediate cells in the stria vascularis are preferentially affected by cisplatin
2083 exposure. Additionally, scRNA-seq data revealed specific gene expression changes
2084 associated with chemotherapy-induced ototoxicity. Notably, genes such as Alcam,
2085 Atp1b2, Spp1, and Car12 were downregulated, while Klf10, Cldn3, and Tspan1 were
2086 upregulated in cisplatin-treated stria vascularis. These differentially expressed genes
2087 present potential novel therapeutic targets to mitigate ototoxicity caused by
2088 chemotherapy (Taukulis et al., 2021). Zhang et al. investigate the immunomodulatory
2089 mechanisms of dihydroartemisinin using scRNA-seq in combination with cellular and
2090 biochemical methods (Zhang et al., 2022f). Their research revealed that
2091 dihydroartemisinin beneficially regulated immune cell heterogeneity and splenic
2092 immune cell homeostasis by activating the SOD3-JNK-AP-1 pathway to treat
2093 autoimmune diseases. Understanding the mechanisms of drug action is vital for
2094 optimizing therapeutic efficacy and minimizing adverse effects.

2095

2096 *Patient stratification*

2097 ScRNA-seq is a useful ally of personalised medicine, which seeks to customise
 2098 treatment plans based on unique patient features. This technology contributes to
 2099 patient stratification by profiling the transcriptomes of individual cells, allowing for
 2100 the identification of markers relevant to disease prognosis or therapeutic response.

2101 In the context of infant acute lymphoblastic leukemia (iALL), where relapse
 2102 occurrence is often fatal, scRNA-seq has shown promise (Pieters et al., 2019). Using
 2103 diagnostic samples from patients with MLL-rearranged infant acute lymphoblastic
 2104 leukemia (MLL-r iALL), Candelli et al. employed scRNA-seq. By measuring the
 2105 percentage of cells found to be either sensitive or resistant to therapy, the researchers
 2106 were able to forecast when MLL-r iALL would relapse. This approach outperformed
 2107 current risk stratification schemes, showcasing the potential of scRNA-seq in refining
 2108 prognostic markers for better treatment outcomes (Candelli et al., 2022).

2109 Our knowledge of cellular heterogeneity, disease processes, and treatment
 2110 responses at the single-cell level has been completely transformed by the use of
 2111 scRNA-seq in drug discovery and development. Its contributions span the
 2112 identification of therapeutic targets, improvement of drug screening and optimization,
 2113 elucidation of mechanisms of action, and facilitation of patient stratification. The
 2114 integration of scRNA-seq in drug discovery holds significant promise for developing
 2115 more effective and personalized therapies, ultimately leading to improved patient
 2116 outcomes.

2117

2118 *Summary*

2119 The evolution of single-cell transcriptomic atlases through advancements in
 2120 scRNA-seq technology has provided unprecedented resolution, offering insights into
 2121 complex cellular events and enhancing our understanding of cell composition and
 2122 interactions across humans, model animals, and plants. This review underscores the
 2123 progress in different aspects of scRNA-seq technology, emphasizing distinct features
 2124 and strengths in various areas. It is crucial to recognize that each single-cell
 2125 sequencing method has its own advantages and limitations. Ongoing developments in
 2126 this field aim to design improved methods that enhance robustness and coverage,
 2127 allowing for comprehensive detection of cellular composition at multiple levels and
 2128 the depiction of cell landscapes within different species. The expectation is that future
 2129 innovations in scRNA-seq technologies will contribute significantly to the
 2130 advancement of biological and clinical medicine, offering powerful tools for in-depth
 2131 exploration and understanding of cellular dynamics.

2132

2133 **Chapter 2 Single-cell whole-genome sequencing**

2134

2135 The microscopization of life science research proves that cell population-based
 2136 methods may not be suitable for certain areas of study, such as tumor heterogeneity
 2137 and early embryonic development. In response, the introduction of single-cell
 2138 transcriptome sequencing technology in 2009 marked a significant advancement
 2139 (Tang et al., 2009). Building on this, Navin et al. introduced single-cell whole

2140 genome sequencing technology (scWGS) in 2011 by combining whole genome
 2141 amplification (WGA) with high-throughput sequencing (Navin et al., 2011). This
 2142 innovative approach addresses challenges related to obtaining information about
 2143 heterogeneity between different cells in tissue samples and enables the study of
 2144 individual cells when conventional sequencing might be impractical due to small
 2145 sample sizes.

2146 By sequencing DNA at the single-cell level, scWGS provides a new dimension
 2147 for studying the behavior and mechanisms of individual cells. Applications of scWGS
 2148 have expanded across various research fields, including neuroscience, germline
 2149 evolution, organogenesis, oncology, clinical diagnosis, immunology, microbiology,
 2150 embryo development, and prenatal genetic diagnosis. Recognizing its potential,
 2151 scWGS was highlighted as one of the most anticipated technologies in 2013 by the
 2152 journal *Nature Methods*.

2153 The development of scWGS has indeed opened up avenues for researchers to
 2154 delve into inter-cell heterogeneity at the single-cell level, exploring various aspects
 2155 such as single-nucleotide variants, short insertions or deletions, and copy number
 2156 variants. This technology has proven particularly valuable for studying the genomes
 2157 of rare cells that hold biological or clinical significance, including circulating tumor
 2158 cells and cells used in third-generation in vitro fertilization preimplantation genetic
 2159 diagnosis/screening. The scWGS process typically involves three main steps:
 2160 single-cell isolation, single-cell whole genome amplification (scWGA), and the
 2161 sequencing and analysis of the amplified products. The critical challenge in this
 2162 process is to amplify the genome of a single cell effectively, obtaining sufficient
 2163 material for downstream analyses while minimizing artifacts such as amplification
 2164 bias, genome loss, mutations, and chimeras. Addressing these challenges is essential
 2165 to ensure the accuracy and reliability of the genetic information obtained from single
 2166 cells.

2167 In this section, we first focus on the advancements in scWGA technology.
 2168 Subsequently, we provide a detailed introduction to several prominent scWGA
 2169 chemistries, elucidating their crucial biochemical reaction strategies. Our focus then
 2170 shifts to high-throughput scWGS methods, which enable the parallel sequencing of
 2171 tumor cell genomes on a massive scale. This approach opens up opportunities to
 2172 significantly broaden the scope of intratumoral characterization. Key milestones in the
 2173 development of scWGA and high-throughput scWGS technologies are visually
 2174 represented in Figure 5. Lastly, we offer a summary of the most recent practical
 2175 breakthroughs in scWGS within the field of biomedicine. This overview outlines a
 2176 vision for applying single-cell genomic sequencing in clinical research, highlighting
 2177 its potential impact on advancing our understanding of biological processes and
 2178 disease mechanisms.

2179

2180 *ScWGA methods*

2181 Given the limited DNA content in a single cell (approximately 6 pg/cell), which falls
 2182 short of the detection requirements of sequencers, it is imperative to amplify the trace
 2183 amounts of whole-genome DNA in single cells before sequencing. This amplification

process aims to generate a complete genome with high coverage, ensuring accurate and comprehensive sequencing results in subsequent high-throughput sequencing. Over time, major changes have occurred in WGA technology to meet these demands. Notable methods include degenerate oligonucleotide-primed polymerase chain reaction (DOP-PCR) (Telenius et al., 1992), multiple displacement amplification (MDA) (Dean et al., 2001), and multiple annealing and looping-based amplification cycles (MALBAC) (Zong et al., 2012). Subsequent innovations, such as linear amplification via transposon insertion (LIANTI) (Chen et al., 2017a), single-stranded sequencing using microfluidic reactors (SISSOR), primary template-directed amplification (PTA) (Gonzalez-Pena et al., 2021), and multiplexed end-tagging amplification of complementary strands (META-CS) (Xing et al., 2021) have further expanded the WGA toolkit. Table 1 provides an overview of the general characteristics of these methods. In the following sections, we conduct a review of several major WGA methods, focusing on their coverage, uniformity, and accuracy. In short, there is no obvious winner in amplification performance, and each strategy has advantages according to the parameter that matters.

2200

2201 *PCR-based amplification*

2202 PCR-based WGA methods come in various forms, including Primer Extension
 2203 Preamplification PCR (PEP-PCR) (Zhang et al., 1992), DOP-PCR (Telenius et al.,
 2204 1992), tagged random primer PCR (T-PCR) (Grothues et al., 1993) and
 2205 ligation-mediated PCR (LM-PCR) (Klein et al., 1999), etc. These techniques were
 2206 essential in obtaining the amplification of genomic DNA from single cells, meeting
 2207 particular research objectives, and acting as models for scWGA technology. One of
 2208 the first scWGA techniques used was PEP-PCR, which was later superseded by the
 2209 more popular DOP-PCR. These techniques have led to the development of
 2210 commercial kits, such as the PicoPLEX WGA Kit (Rubicon Genomics, USA) and the
 2211 GenomePlex Single Cell Whole Genome Amplification Kit (Sigma-Aldrich, USA).

2212 DOP-PCR operates on the principle of utilizing a partially random primer for a
 2213 two-step PCR amplification of template genomic DNA (Telenius et al., 1992). The
 2214 degenerate primers consist of a random six base sequence in the middle flanked by
 2215 fixed sequences at each end (5' CGACTCGAGNNNNNNATGTGG 3'). The short
 2216 ATGTGG sequence at the 3' end of the primer has an extremely high distribution
 2217 frequency in genomic DNA, guiding the initial low-temperature annealing step and
 2218 determining the starting site of amplification with a bias towards specific sequences.
 2219 The middle six degenerate bases create 46 different sequences, and during annealing,
 2220 one or more of these degenerate bases, along with the 3' end specific bases,
 2221 simultaneously bind to the template DNA, enhancing the primer's binding efficiency.
 2222 The PCR amplification occurs in two steps: the first few cycles (3~5) involve
 2223 low-temperature annealing (e.g., 30°C), followed by strand extension at an elevated
 2224 temperature. In the second step, the products from the first step undergo further
 2225 amplification using a primer targeting the 5' fixed sequence at a higher annealing
 2226 temperature (62°C).

2227 The efficiency of DOP-PCR amplification relies on primer concentration and

polymerase activity. At a low annealing temperature, the primers can bind to multiple genomic loci, resulting in amplification products that cover nearly the entire genome. DOP-PCR stands out as a representative method in PCR-based WGA, finding application in amplifying minute amounts of human genomic DNA in single cells for analyses related to tumor heterogeneity, assessment of copy number variations, and detection of aneuploidies (Navin et al., 2011; McConnell et al., 2013; Knouse et al., 2014). However, DOP-PCR often generates low genome coverage (typically less than 10%) (Navin et al., 2011), a characteristic associated with the exponential nature of PCR amplification. Additionally, the PCR amplification reaction has a high base mismatch rate, rendering this amplification technology less suitable for the detection of single-nucleotide variations due to the elevated false positive rate.

2239

2240 *Isothermal amplification*

MDA is the most representative method among isothermal scWGA methods, initially developed by Dean et al. in 2001 (Dean et al., 2001). Operating under isothermal conditions, MDA employs a 6 base pair random primer that randomly anneals to the genome, initiating a strand displacement amplification reaction catalyzed by phi29 DNA polymerase with robust strand displacement activity. The single-stranded sequence produced through displacement can be extended randomly by annealing with random primers, resulting in the formation of multibranched amplification structures. Due to the potent DNA synthesis ability of phi29 DNA polymerase, the synthesized DNA fragments are typically 50~100 kb in length. Additionally, the high replication fidelity of phi29 DNA polymerase, characterized by an error rate of about one nucleotide per 10^8 , owing to its 3'→5' exonuclease and proofreading activities, makes MDA suitable for accurate single nucleotide variation (SNV) calling. This feature has led to its application in single-cell genome lineage tracing (Lodato et al., 2015). Furthermore, MDA provides significantly higher genome coverage compared to initial PCR-based methods. However, a drawback of MDA is its exponential amplification process, similar to DOP-PCR, which introduces sequence-dependent bias and hinders coverage uniformity. It's worth noting that the sequence-dependent bias of MDA is not consistently reproducible across the genome from one cell to another, making copy number variation (CNV) measurements noisy and normalization less effective.

To address amplification bias and enhance uniformity and coverage, various improved isothermal amplification technologies, including emulsion MDA (eMDA) (Fu et al., 2015), digital droplet MDA (ddMDA) (Sidore et al., 2016), TruePrime (Picher Á et al., 2016), SISSOR (Chu et al., 2017) and PTA (Gonzalez-Pena et al., 2021) have been developed based on MDA technology. Both eMDA and ddMDA involve dispersing the amplification process into millions of small droplets, aiming to improve amplification uniformity and correct bias. The TruePrime technique substitutes the N6 primer in the MDA method with a unique DNA primase called TthPrimPol to enhance amplification uniformity. SISSOR enhances sequencing accuracy by randomly distributing megabase-sized single-stranded DNA fragments from homologous chromosome pairs into numerous nanoliter compartments for

enzymatic amplification within a microfluidic device. By adding exonuclease-resistant terminators to the reaction, the PTA technique generates smaller double-stranded amplification products that perform limited subsequent amplification. This causes the reaction to change from an exponential to a quasilinear process, increasing the amount of amplification that comes from the primary template and enhancing the coverage and uniformity of genome amplification. Currently, MDA-based products are well-established such the REPLI-g Single Cell Kit from Qiagen.

2280

2281 *MALBAC*

Reported in 2012 by Zong et al. (Zong et al., 2012), MALBAC (Multiple Annealing and Looping-Based Amplification Cycles) is a scWGA method designed to mitigate bias associated with nonlinear amplification. MALBAC primers feature a common 27 nucleotide sequence at the 5' end and 8 random nucleotides at the 3' end. The amplification process begins with the hybridization of these 8 random nucleotides to the genomic DNA template at 0°C. Then, as the temperature is increased to 65°C, DNA polymerases with strand-displacement activity are used to create semiamplicons of varying lengths (0.5 to 1.5 kb). At 94°C, the semiamplicons are then denatured from the template. The semiamplicons are amplified further to produce entire amplicons with complementary ends, forming hairpins when the temperature is lowered (to 58°C). Full amplicons form loops that may impede further amplification and cross-hybridization. Following five rounds of preamplification, oligonucleotides containing the common 27 nucleotide sequence are used as primers in PCR to exponentially amp up the entire amplicons. This process yields the micrograms of DNA required for next-generation sequencing. Additionally, commercial kits based on the MALBAC method have been created, such as the MALBAC single-cell WGA test (Qiagen).

2299

MALBAC stands out as more than just a combination of DOP-PCR and MDA; it is fundamentally distinct due to its quasilinear amplification, which circumvents sequence-dependent bias associated with exponential amplification, thereby enhancing amplification uniformity. In the initial phase of MALBAC technology, multiple displacement reactions are employed to achieve a whole genome coverage of the amplification product up to 93%. MALBAC boasts both extensive coverage and uniform amplification, making it suitable for the genome-wide detection of SNPs and CNVs in a single cell (Zong et al., 2012; Hou et al., 2013). The technology's implementation has contributed to the advancement of clinical assisted reproductive technology (Yao et al., 2018). Furthermore, MALBAC demonstrates a low false negative rate in SNV detection. However, when compared to MDA technology, MALBAC does exhibit a higher false positive rate in SNV detection due to the lower fidelity of the current DNA polymerase used in comparison to the phi29 polymerase.

2312

2313 *LIANTI*

In contrast to previous WGA methods, such as the exponential PCR reaction with degenerate priming in DOP-PCR (Telenius et al., 1992), the strand-displacing DNA

polymerase-driven exponential amplification of single-stranded DNA in MDA (Dean et al., 2001), and the quasilinear amplification in MALBAC (Zong et al., 2012) through looping-based amplicon protection followed by PCR all of which involve nonspecific priming and exponential amplification leading to bias and errors, a novel scWGA method was developed by Xie et al.'s team in 2017, known as Linear Amplification via Transposon Insertion (LIANTI) (Chen et al., 2017a). LIANTI uses the Tn5 transposase technology to sustain linear amplification during the whole genome amplification process. This method randomly slices genomic DNA, and then fills in the cut locations with a predetermined sequence. By utilizing the Tn5 transposase property, LIANTI inserts the T7 promoter into genomic DNA. The genomic DNA fragments labelled with T7 promoters are then linearly amplified into thousands of copies of RNAs using T7 RNA polymerase for IVT. RT and second-strand synthesis come next, producing double-stranded LIANTI amplicons that are prepared for DNA library construction. In comparison to earlier techniques (DOP-PCR, MDA, and MALBAC), LIANTI exhibits superior sensitivity and accuracy in identifying CNVs and SNVs, respectively. Additionally, LIANTI outperforms other WGA techniques (DOP-PCR, MDA, and MALBAC) with a 97% genome coverage and a 17% allele dropout rate.

2334

2335 *META-CS*

True SNV must be located at the same position on both DNA strands, while polymerase mistakes and DNA damage often happen at random on one of the two strands. Therefore, sequencing both complementary strands of double-stranded DNA (dsDNA) is essential to reduce false positives on single strands and improve accuracy. Multiplexed End-Tagging Amplification of Complementary Strands (META-CS), a revolutionary scWGA technique, was presented by Xie and colleagues in 2021 (Xing et al., 2021). Because of DNA complementarity, META-CS can clearly identify and amplify the two DNA strands in a one-tube reaction, reducing almost all false positives. De novo SNVs can be reliably identified using this method from a single cell.

META-CS is built upon the previously reported multiplexed end-tagging amplification method developed by this research team. Initially, a combination of 16 unique transposon sequences is mixed with Tn5 transposase in an equal molar ratio to create transposon complexes. These complexes randomly cut the genomic DNA from a single cell. Subsequently, genomic DNA fragments tagged by two random transposon sequences undergo denaturation through heating, releasing two single strands. To obtain strand-specific labelling, these strands are subsequently preamplification using two consecutive polymerase extension processes. After each polymerase extension reaction, exonuclease I is employed to eliminate excess primers. The products, separately amplified from the sense and antisense strands of the original DNA, can be differentiated by mapping them to the reference genome. As a result, SNVs are determined with information from both strands, significantly improving accuracy.

2359

2360 *High-throughput scWGS methods*

2361 Since the introduction of the first scWGS method for profiling copy numbers in
 2362 human tissues (Navin et al., 2011), the field of single-cell genomics has witnessed
 2363 rapid progress over the past decade. Early techniques were confined to profiling a
 2364 small number of cells at a time and were based on WGA chemicals (Telenius et al.,
 2365 1992; Dean et al., 2001; Chen et al., 2017a). However, the study of tumor cell genome
 2366 evolution and reconstruction often requires the analysis of mutation signatures in a
 2367 large number of single-cell tumor genomes simultaneously. Low-throughput scWGS
 2368 methods pose challenges due to their labor-intensive, costly, time-consuming nature,
 2369 and limited efficiency. Recent developments in combinatorial indexing, nanowell, and
 2370 microdroplet techniques have greatly increased cell throughput and decreased costs
 2371 (Vitak et al., 2017; Laks et al., 2019; Yin et al., 2019c; Minussi et al., 2021). In the
 2372 following sections, we will delve into a detailed discussion and comparison of
 2373 high-throughput scWGS methods based on different strategies. The characteristics of
 2374 several key high-throughput scWGS methods are summarized in Table S5.

2375

2376 *Microfluidic-based high-throughput scWGS methods*

2377 *Direct library preparation method (DLP)*

2378 A direct DNA transposition single-cell library production (DLP) approach was
 2379 presented by Zahn et al. in 2017 (Zahn et al., 2017). This method creates indexed
 2380 libraries straight from single cells. The DLP method utilizes a specially designed
 2381 microfluidic device to capture and lyse single cells. Tn5 transposome complexes
 2382 randomly fragment the genomic DNA of a single cell, tagging each fragment at the 5'
 2383 end with a distinct adaptor sequence. The index barcodes and sequencing adaptors are
 2384 then added to both ends of the fragmented DNA inserts using eleven PCR cycles.
 2385 Following indexing, the libraries are combined for multiplexed sequencing.

2386 Prior to building libraries, early single-cell techniques used WGA to capture
 2387 entire genomes (Navin et al., 2011; Gawad et al., 2014; Wang et al., 2014b). However,
 2388 preamplification introduces amplification biases and reduces coverage uniformity,
 2389 hindering the detection of CNVs (Wang et al., 2012; Zong et al., 2012; Macaulay and
 2390 Voet, 2014). DLP is an example of the first direct production of single-cell libraries
 2391 without preamplification using fragmentation. This method produces genomes with
 2392 high uniform coverage and enables multiplexing of many cells, which makes it
 2393 appropriate for high-throughput and reasonably priced CNV detection. Compared to
 2394 DOP-PCR, DLP is more cost-effective (approximately \$0.50 per cell versus \$15 per
 2395 cell) and time-efficient (2.5 hours versus 3 days). However, the use of microfluidic
 2396 devices limits the throughput of cells and necessitates a certain size of cells because
 2397 very small cells may slip through traps unless the devices are specifically made for
 2398 that type of cell. Large cells can also clog channels. Thus, additional optimization is
 2399 still required.

2400

2401 *Microfluidic droplet method*

2402 Increasing the throughput of cellular sequencing faces challenges due to limitations in
 2403 single-cell partitioning methods, difficulties in amplifying genomic DNA from single

2404 cells, and the complexity of enzymology steps for library preparation (Lan et al., 2017;
 2405 Vitak et al., 2017). In response to these challenges, Andor et al. proposed a solution to
 2406 enable large-scale scWGS (Andor et al., 2020). Their method makes use of a
 2407 two-stage microfluidic droplet-based technique to automatically generate scWGS
 2408 libraries with a high cell number. Similar to previous single-cell transcriptome
 2409 investigations, microfluidic droplets are loaded with a barcoded hydrogel bead that
 2410 labels DNA. To create cell beads (CBs), individual cells are first encased in a
 2411 hydrogel matrix. Reagents for cell lysis and protein digestion are then added to these
 2412 CBs to lyse and unpackage DNA. To create cell bead-gel bead (CBGB) emulsion, a
 2413 second microfluidic chip is utilized in which a gel bead (GB) is functionalized with
 2414 millions of copies of a distinct droplet-identifying barcode and co-encapsulated with
 2415 the hydrogel CB and enzymatic reaction mix. CBGB dissolves to release contents
 2416 after encapsulation. Genomic DNA fragments labelled with a sequencing adaptor and
 2417 a barcode sequence are obtained by a two-step isothermal incubation process. After
 2418 breaking and purifying the emulsion, the library is ready for Illumina sequencing.
 2419 This microfluidic droplet-based cellular isolation technology can isolate tens of
 2420 thousands of cells in a single experiment at a throughput that surpasses that of
 2421 conventional DLP techniques.

2422

2423 *Nanowell-based high-throughput scWGS methods*

2424 *High-throughput direct transposition scWGS method (DLP+)*

2425 When compared to preamplification-based techniques, the earlier DLP method, which
 2426 made use of microfluidic devices, effectively decreased biases (Zahn et al., 2017).
 2427 Despite the good performance of microfluidic-based DLP analysis, the usage of
 2428 customised microfluidic devices limits cell size and presents obstacles to their general
 2429 adoption and scalability. Additionally, some droplet-based methods face similar
 2430 constraints on cell size (Andor et al., 2020). In response to these problems, Emma
 2431 Laks and colleagues developed the DLP+ platform, a higher-throughput direct
 2432 transposition scWGS system built on commercially available “off the shelf” picoliter
 2433 volume piezo-dispensing technology and commodity high-density nanowell arrays
 2434 (Laks et al., 2019). In the DLP+ method, single cells are isolated through limiting
 2435 dilution and dispensed into nanowell arrays. To achieve an almost flawless single-cell
 2436 isolation rate, chosen cells are deposited into reaction chambers selectively using
 2437 spotting software, which locates them inside the dispensing nozzle. A 10x inverted
 2438 fluorescence microscope scans each nanowell chip to verify single-cell occupancy and
 2439 gather data on the cell state. Before the library preparation reagents are spotted,
 2440 imaging takes place, enabling the exclusion of doublets, empty wells, or contaminated
 2441 cells from the procedure. Adding reagents, spinning, sealing, and heating the chip are
 2442 the procedures involved in creating DLP+ libraries from unamplified single cells.
 2443 Using normal Illumina procedures and HiSeq equipment, the resulting libraries are
 2444 pooled during recovery and sequenced at the required coverage depth.

2445 With its transparent dispensing nozzle and inbuilt camera, DLP+ provides a
 2446 unique benefit that makes it possible to take high-resolution microscope images of
 2447 objects prior to dispensation. By actively selecting individual cells, this function helps

2448 to prevent the sequencing of detritus or doublet cells. Furthermore, by methodically
 2449 modifying a number of variables, including cell lysis volume and buffer type,
 2450 transposase (Tn5) concentration, post-indexing PCR cycles, and cell lysis/DNA
 2451 solubilization duration, DLP+ has refined the physical reaction determinants for
 2452 producing high-quality libraries. The DLP approach is the foundation for these
 2453 optimizations (Zahn et al., 2017). Moreover, DLP+ has shown genome coverage
 2454 uniformity that is comparable to that of the microfluidic-based DLP approach, but at a
 2455 throughput that is significantly higher, scaling from hundreds to tens of thousands of
 2456 cells per experiment across a range of tissue types.

2457

2458 *Archival nanowell sequencing method (Arc-well)*

2459 The previously developed high-throughput scWGS methods, including DLP (Laks et
 2460 al., 2019), SCI-seq (Vitak et al., 2017), and ACT (Minussi et al., 2021), have a
 2461 common limitation—they require fresh or snap-frozen tissue samples, rendering them
 2462 unsuitable for the analysis of archival formalin-fixed paraffin-embedded (FFPE)
 2463 tissue samples. Addressing this challenge, a novel method called Arc-well (archival
 2464 nanowell sequencing) was introduced by Wang et al. in 2023 (Wang et al., 2023a). To
 2465 perform Arc-well, FFPE blocks are sectioned and deparaffinized to produce
 2466 single-nucleus suspensions, which are subsequently used for FACS sorting. After the
 2467 sorted nuclei are distributed into a 5,184-well nanowell chip, the nanowells can be
 2468 imaged to select single cells and prevent doublets, deteriorated nuclei, and empty
 2469 wells. Then, a five-step equal volume dispensing step is performed by using the
 2470 ICELL8 cx system (Takara Bio), which is used to dispense downstream reagents into
 2471 nanowell chips. First, lysis reagents are dispensed to lysis selected nuclei and release
 2472 the genomic DNA. Next, reagents for labeling reaction (Tn5 transposome) and Tn5
 2473 inactivation are dispensed. Furthermore, by depositing dual indices (72 × 72
 2474 combinations) and amplifying the PCR result, every nanowell is given a distinct
 2475 barcode combination. The barcoded libraries are then pooled and sequenced on the
 2476 Illumina platforms.

2477 The acoustic cell fragmentation (ACT) technique was first presented by the
 2478 researchers in 2021. This technique made use of acoustic liquid transfer (ALT)
 2479 technology, direct fragmentation of genomic DNA, and FACS of single nuclei to
 2480 enable high-throughput single-cell DNA sequencing at single-molecule resolution
 2481 (Minussi et al., 2021). When comparing the ACT method with Arc-well, it was found
 2482 that Arc-well exhibited higher throughput (1,900~2,600 cells per experiment), lower
 2483 reagent costs, and reduced technical variability. Importantly, Arc-well demonstrated
 2484 the capability to amplify degraded DNA fragments commonly found in archival FFPE
 2485 tissues, making it compatible with such tissue samples.

2486

2487 *Combinatorial indexing-based high-throughput scWGS methods*

2488 *SCI-seq & SCI-L3-WGS*

2489 FACS is used in the SCI-seq method to sort individual cells into 96-well plates (Vitak
 2490 et al., 2017). Subsequently, genomic DNA from a single cell undergoes random
 2491 fragmentation by Tn5 transposase, and each resulting fragment is tagged with index 1

2492 and an adaptor. The introduction of index 2 is achieved through a PCR reaction.
2493 Ultimately, the distinct libraries are combined for sequencing. Nucleosome depletion
2494 is used in a combinatorial indexing procedure by SCI-seq, which makes it possible to
2495 produce thousands of single-cell genome sequencing libraries at once. This approach
2496 has also the benefit of not requiring specialized microfluidics equipment or droplet
2497 emulsification procedures, in addition to its high throughput. However, it is noted that
2498 SCI-seq technology introduces a certain bias during the PCR amplification process.

2499 To address amplification bias, Yin et al. introduced sci-L3 (Yin et al., 2019c), a
2500 method that integrates combinatorial indexing with linear amplification. With the help
2501 of a 3-level indexing technique, sci-L3-WGS considerably increases LIANTI's
2502 throughput, allowing it to sequence at least thousands or even millions of cells per
2503 experiment while minimizing amplification biases. The sci-L3-WGS process is
2504 delineated into three key steps: (1) Tn5 transposase randomly cleaves genomic DNA
2505 from a single cell and attaches barcode 1 to each fragment. (2) A second set of
2506 barcodes is ligated to the ends of DNA fragments, along with a T7 promoter
2507 positioned outside both barcodes. (3) The introduced T7 promoter initiates IVT,
2508 followed by RT and second-strand synthesis. A third set of barcodes and UMIs are
2509 introduced during second-strand synthesis. Duplex DNA molecules can be prepared
2510 in accordance with conventional library preparation procedures. Each molecule
2511 contains three barcodes that identify the cell of origin. The sci-L3 strategy has a
2512 number of benefits over current methods and any straightforward combination of
2513 SCI-seq (Vitak et al., 2017) and LIANTI (Yin et al., 2019c). First off, using IVT, it
2514 accomplishes the same linear amplification as LIANTI. Second, because it uses three
2515 rounds of barcoding, its theoretical throughput surpasses one million cells per
2516 experiment at a cheap cost of library preparation (Cao et al., 2019). Thirdly, sci-L3 is
2517 a flexible strategy for linear amplification combined with high-throughput cellular
2518 indexing; it can be used for other single-cell sequencing analysis besides scWGS,
2519 such as single-cell RNA/DNA co-assays.

2520

2521 *Applications of scWGS in biomedicine*

2522 With its ability to reveal differences in single-cell genomic architecture, scWGS
2523 technology is a potent tool that is used in many different fields, including tumour
2524 biology, somatic mutation and mosaicism, organismal development, germ cell
2525 mutation and development, fertility, and microbial research. It has become a major
2526 area of study in the life sciences. The applications of this technology in the fields of
2527 fertility and tumour biology will be the focus of the discussion that follows.

2528

2529 *Tumor biology*

2530 Cancer, a multifaceted and diverse disease characterized by genomic instability and
2531 the accrual of somatic mutations, poses a significant challenge for comprehensive
2532 understanding. Traditional bulk sequencing methods have offered valuable insights
2533 into the genomic makeup of cancer; however, they often overlook the inherent
2534 heterogeneity within tumors, resulting in an incomplete portrayal of the disease. The
2535 advent of scWGS has proven instrumental in overcoming this limitation. By enabling

2536 the analysis of individual cancer cells at a single-molecule level, scWGS has
 2537 exhibited significant potential in various facets of cancer research, such as elucidating
 2538 intratumoral heterogeneity, interpreting the evolution of clonal processes,
 2539 comprehending invasion and metastasis, investigating circulating tumour cells (CTCs),
 2540 and evaluating treatment outcomes.

2541 A study concentrating on breast malignancies revealed the first investigation of
 2542 intratumoral heterogeneity utilizing scWGS based on DOP-PCR. This study identified
 2543 subclonal lineages inside breast tumours using copy number changes (Navin et al.,
 2544 2011). Subsequent scWGS studies, utilizing diverse cancer types such as ovarian
 2545 (McPherson et al., 2016), bladder (Li et al., 2012), brain (Francis et al., 2014), renal
 2546 (Xu et al., 2012), colorectal (Leung et al., 2017; Liu et al., 2017), liver (Hou et al.,
 2547 2016), lung (Ferronika et al., 2017), and hematological (Gawad et al., 2014; Hughes
 2548 et al., 2014a) cancers, have expanded our understanding of intratumoral heterogeneity
 2549 at the levels of CNVs and SNVs. These investigations have revealed a correlation
 2550 between tumor subtype and subclonal diversity in specific cases. For example, Baslan
 2551 et al. conducted a comprehensive analysis of 2,086 breast cell genomes from 16 breast
 2552 cancer samples using a DOP-PCR-based sequencing method. They observed that
 2553 estrogen receptor-negative breast cancers exhibit higher subclonal diversity compared
 2554 to estrogen receptor-positive breast cancers (Baslan et al., 2020).

2555 Phylogenetic analyses based on intratumoral heterogeneity profiles obtained
 2556 through single-cell DNA sequencing (scDNA-seq) provide valuable insights into
 2557 identifying driver mutations-genetic alterations that play a significant role in cancer
 2558 development and progression. Analyzing the genomes of individual cancer cells
 2559 allows researchers to pinpoint specific mutations driving tumor growth, offering
 2560 crucial information for the development of targeted therapies that address these driver
 2561 mutations. In a work by Wang et al., hundreds of breast cells were profiled using a
 2562 combination of targeted duplex single-molecule sequencing and scWGS (Wang et al.,
 2563 2014b). In two individuals with breast cancer, the researchers looked into mutational
 2564 evolution and clonal diversity. Their research showed that SNVs gradually evolved,
 2565 resulting in a high degree of clonal diversity. On the other hand, aneuploid
 2566 rearrangements happened early in the genesis of tumours and stayed very stable
 2567 during clonal growth. The investigation discovered many nonsynonymous mutations
 2568 in genes linked to cancer, such as PIK3CA, CASP3, FBN2, and PPP2R5E, in a
 2569 sample of invasive ductal carcinoma that was positive for oestrogen receptors.
 2570 Interestingly, it is known that the most frequent driver mutation in luminal A breast
 2571 tumours is PIK3CA (Ellis et al., 2012; Network, 2012).

2572 CTCs, originating from primary tumors and entering the peripheral blood, have
 2573 the potential to contribute to metastasis. ScWGS of CTCs presents a promising
 2574 approach for noninvasive sampling of tumors, offering insights into noninvasive
 2575 prognosis or even diagnosis. In a study by Riebensahm et al., scWGS was employed
 2576 to analyze the mutation characteristics of genes in CTCs from breast cancer brain
 2577 metastasis patients (Riebensahm et al., 2019). The study identified mutated genes
 2578 such as TP53, ARID1A, CDH1, and TTN, with ARID1A, involved in chromatin
 2579 remodeling, highlighted as a potential druggable target. The MALBAC approach was

employed by Ni et al. to examine the genomes of individual CTCs obtained from patients with lung cancer (Ni et al., 2013). The analysis revealed the presence of insertions/deletions (indels) and SNVs that are linked to cancer in the CTC exomes. This mutation information provided potential clinical guidance for personalized therapy. Additionally, CTCs have been utilized for noninvasive monitoring of treatment response (Dago et al., 2014).

In summary, scWGS stands as a groundbreaking technology in cancer research, providing a comprehensive understanding of the genomic landscape of individual cancer cells. By uncovering clonal evolution, identifying driver gene mutations, tracking chromosomal abnormalities, studying the tumor microenvironment, and detecting minimal residual disease, scWGS offers valuable insights for cancer diagnosis, prognosis, and targeted therapy. As scWGS technology continues to develop, it holds great promise for advancing personalized cancer medicine.

2593

Fertility

Preimplantation genetic diagnosis (PGD) and preimplantation genomic screening (PGS) for embryos created through in vitro fertilization (IVF) are two clinical uses for scWGS. This helps prevent the inheritance of harmful mutations and chromosomal abnormalities by enabling a thorough study of chromosomes. For this, a variety of genome analysis systems are used, including multiplex quantitative PCR, comparative genomic hybridization (CGH) arrays, and single-nucleotide polymorphism (SNP) arrays (Treff et al., 2012; Rubio et al., 2013; Tobler et al., 2014). By enhancing traditional techniques for examining embryo biopsies, scWGS technologies allow for the simultaneous identification of mutations and aneuploidy throughout the genome (Treff et al., 2013; Wells et al., 2014; Kumar et al., 2015). The rapid advancement of high-throughput sequencing methods has further decreased expenses and improved the accuracy and resolution of PGD/PGS at the chromosomal level. This approach holds promise for enhancing the accuracy of selecting healthy embryos during IVF procedures, improving the success rates of assisted reproduction, and reducing the risk of genetic disorders in newborns. Below we describe several application examples of scWGS in PGD/PGS.

The use of scWGS in PGS and PGD during IVF has been shown in a number of research: Wells et al. utilized DOP-PCR-based WGA to perform scWGA on the first polar body, successfully detecting chromosomal abnormalities in embryos using CGH technology (Wells et al., 2002). Daina et al. conducted monogenic analysis on fourteen embryos for a family affected by Lynch syndrome, achieving successful double-factor PGD using the MDA method and leading to the birth of two healthy children (Daina et al., 2013). Hou et al. employed MALBAC-based sequencing technology to analyze the genomes of single human oocytes from eight healthy donors. They demonstrated how accurate and economical selection of normal fertilized eggs for embryo transfer is made possible by MALBAC-based preimplantation genomic screening in vitro fertilization (Hou et al., 2013). Huang et al. collected 23 frozen cleavage embryos from three pregnant women donors and performed single-cell CGH, SNP, and MALBAC sequencing for 24-chromosome

aneuploidy analysis. MALBAC sequencing results showed a high concordance rate with CGH and SNP, indicating its application value in PGD/PGS (Huang et al., 2014). Shang et al. extended the application of MALBAC-scWGS to PGD/PGS detection of mitochondrial disorders, demonstrating the versatility of this technology in addressing various genetic conditions (Shang et al., 2018).

ScWGS has revolutionized PGD/PGS detection by enabling the analysis of individual cells within embryos. This powerful technique provides detailed information on chromosomal abnormalities, structural variations, and mutational landscapes. The ability to examine the genomic content of individual cells within embryos enhances the precision of genetic analysis, offering valuable insights for selecting embryos with the highest likelihood of success during IVF. As a result, scWGS has contributed to improving the success rates of IVF procedures.

2636

2637 *Summary*

2638 The progress in single-cell genomics technology has not kept pace with that of
 2639 transcriptomics, mainly due to challenges in achieving even genomic coverage during
 2640 DNA capture. Nonetheless, single-cell genome sequencing has brought about
 2641 significant insights into various previously inaccessible biological questions. This
 2642 technology has found applications in diverse research fields, including somatic
 2643 mutagenesis, understanding genome function, studying organismal development, and
 2644 exploring microbiology. Single-cell genome sequencing shows great potential in
 2645 clinical and translational research and practical applications, especially for the
 2646 oncology and assisted reproduction field.

2647

2648 **Chapter 3 Single-cell epigenome sequencing**

2649

2650 The epigenome of a cell regulates its cell type-specific gene expression. Understanding
 2651 epigenetic variations is crucial to reveal transcriptional mechanisms that determine
 2652 tissue and cellular heterogeneities during development, disease formation, and
 2653 progression. The epigenome involves a variety of precisely regulated epigenetic
 2654 features, such as nucleic acid methylation, chromatin states, nucleosome positions,
 2655 histone modifications (HM), transcription factors (TF) bindings, and high-order
 2656 chromatin structures. These features interact with one another to influence nearby
 2657 genome activity without changing DNA sequences, which further controls cellular
 2658 activities and results in heritable phenotypes. Single-cell epigenome sequencing
 2659 techniques, as well as corresponding computational analysis methods, have been
 2660 developed and widely used in many research areas, especially in cancer immunology,
 2661 embryonic development, and neurobiology. In this chapter, we survey the recent
 2662 advances in sequencing techniques and computational tools developed for single-cell
 2663 epigenome data analysis.

2664

2665 *Techniques for sequencing the single-cell epigenome*

2666 *Methylation*

2667 Methylation is a type of epigenetic modification that adds methyl groups (CH₃) to

nucleic acids. In vertebrates, DNA is mostly methylated at the carbon atom occupying the fifth position of the cytosine ring (5mC). The majority of cytosine methylation generally occurs in the context of CpG dinucleotides, which usually group in CpG-dense regions called CpG islands (CGIs). These regions show high associations with gene promoters, resulting in methylation-regulated gene expression in a *cis* manner. There are also other DNA modifications, such as 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC), which are intermediate products of DNA demethylation and play critical roles in many biological processes. Based on whether they use the bisulfite conversion or not, two mainstream types of sequencing methods have been adopted for profiling DNA methylation at the single-cell level. Furthermore, N6-methyladenosine (m^6A) is also an abundant RNA methylation that affects RNA regulation and cellular functions. Another technique for profiling m^6A from RNA at the single-cell level has been developed as well.

Over the past decades, bisulfite sequencing has become the gold standard for profiling genome-wide DNA methylation. With the sodium bisulfite treatment, the unmethylated cytosines are deaminated to uracil while the methylated cytosines remain unchanged. In the following PCR amplification and sequencing, the unmethylated cytosines are then read as thymine, while the methylated cytosine is still read as cytosine. The efficiency of bisulfite treatment reaches about 95%, and the readout of bisulfite-based sequencing methods achieves single base-pair resolution, which enables them to become the dominant methods. Whole-genome bisulfite-sequencing (WGBS) (Cokus et al., 2008) can cover almost all the CpG sites of the whole genome, but the requirement for very deep sequencing makes it costly. Then, reduced representation bisulfite sequencing (RRBS) (Gu et al., 2010) has been developed as a cost-efficient method. It utilizes restriction enzyme (MspI) digestion and size fractionation to enrich CpG-dense regions so that it reduces the complexity and size of the sequence library.

To overcome the massive loss of DNA when detecting methylation in single cells, the single-cell RRBS (scRRBS) (Guo et al., 2013) protocol was introduced. It integrated all of the experimental processes in a single-tube reaction without any purification steps before the bisulfite conversion process, then performed two rounds of PCR amplification and deep sequencing. To remove the PCR amplification bias, unique molecular identifiers (UMIs) were introduced by quantitative RRBS (Q-RRBS) (Wang et al., 2015). To avoid the bisulfite-induced loss of intact sequencing templates and avoid amplification bias, post-bisulfite adaptor tagging (PBAT) was adopted in scBS-seq (Smallwood et al., 2014; Clark et al., 2017) and scPBAT (Kobayashi et al., 2016). scWGBS (Farlik et al., 2015) implements the PBAT but without the requirement for the preamplification step, which is suitable for high-throughput analysis at low sequencing coverage. Genome-wide CpG coverage is not always needed and expensive; therefore, single-cell locus-specific bisulfite sequencing (SLBS) (Gravina et al., 2015) can be a cheaper choice and is able to directly detect epimutations in DNA methylation patterns. With the prosperous single-cell barcode or separation techniques invention for high-throughput sequencing, such as microfluidics device, combinatorial index, and nucleus sequencing, they were adopted in microfluidic diffusion-based reduced representation bisulfite sequencing (MID-RRBS) (Ma et al., 2018), single-cell

2712 combinatorial indexing for methylation analysis (sci-MET) (Mulqueen et al., 2018) and
 2713 single-nucleus methylcytosine sequencing (snmC-seq) (Luo et al., 2017), respectively.
 2714 To enrich more regions where CpG methylation is functionally relevant, including
 2715 promoters, CpG islands, CTCF insulators, and enhancers, a single-cell extended
 2716 representation bisulfite sequencing (scXRBS) (Shareef et al., 2021) was established by
 2717 leveraging an early barcoding step for high sensitivity and sample multiplexing.

2718 A limitation of bisulfite treatment is that unmethylated cytosine, 5fCs, and 5caCs,
 2719 are all converted to uracil and cannot be discriminated from each other, which hampers
 2720 the investigation of DNA demethylation. Single-cell methylase-assisted bisulfite
 2721 sequencing (scMAB-seq) (Wu et al., 2017b) was established to solve this problem by
 2722 pretreatment of the DNA with the CpG methylation enzyme M.SssI, which converts
 2723 only the cytosines to 5mCs, protects Cs but not 5fCs and 5caCs, and enables direct
 2724 detection of 5fCs and 5caCs as uracils.

2725 Besides bisulfite treatment, methylation-sensitive restriction enzymes (MSREs)
 2726 are also widely used in detecting and sequencing DNA methylation. Restriction
 2727 enzyme-based single-cell methylation assay (RSMA) (Kantlehner et al., 2011) is easy
 2728 to implement, but its results are reported by gel electrophoresis, and it is not a
 2729 quantitative method. A similar method, single-cell restriction analysis of methylation
 2730 (SCRAM) (Lorthongpanich et al., 2013), also detects the methylation by MSREs but
 2731 uses real-time quantitative PCR (RT-qPCR) as readout. These two methods both fail to
 2732 distinguish between heterozygous and hemizygous methylated alleles in diploid cells.
 2733 Along with SCRAM and single-cell genotyping by next-generation sequencing (NGS),
 2734 single-cell analysis of genotype, expression, and methylation (sc-GEM) (Cheow et al.,
 2735 2016) allows for a more reliable assessment of methylation status at specific sites.
 2736 Genome-wide CGI methylation sequencing for single cells (scCGI-seq) (Han et al.,
 2737 2017) achieved high single-cell CGI coverage, which extended the use of MSREs from
 2738 a limited number of loci to CGIs at the genome-scale. To allow genome-wide detection
 2739 of 5hmC marks in single cells, the restriction endonuclease AbaSI was used in
 2740 single-cell hydroxymethylation sequencing (scAba-seq) (Mooijman et al., 2016).
 2741 Without using MSREs, Reporter of Genomic Methylation (RGM) (Stelzer et al., 2015)
 2742 adopted a fluorescent reporter system, which allows for visualization and tracing of
 2743 dynamic changes in DNA methylation.

2744 Apart from these two conventional methods, the enzyme conversion-based
 2745 methods have emerged as less damaging alternatives to bisulfite treatment and thus
 2746 have been applied to single-cell analysis. EM-seq identifies 5mC and 5hmC by using
 2747 two sets of enzymatic reactions. The initial reaction involves TET2 and T4-BGT
 2748 converting 5mC and 5hmC into products resistant to deamination by APOBEC3A.
 2749 Subsequently, the second reaction, employs APOBEC3A to deaminate unmodified
 2750 cytosines, transforming them to uracils (Vaisvila et al., 2021). Recently, sciEM
 2751 combined single-cell combinatorial indexing with enzymatic conversion marks
 2752 significant advancement as the first non-bisulfite single-cell DNA methylation
 2753 sequencing method (Chatterton et al., 2023). Similar strategies have also been adopted
 2754 for RNA methylation detection. Global RNA m⁶A profiling reveals its functions in
 2755 gene expression control, physiological processes, and disease states. Deamination

adjacent to RNA modification targets (DART-seq) utilizes a fusion protein consisting of the m6A-binding YTH domain tethered to the cytidine deaminase APOBEC1 (APOBEC1-YTH) to conduct C-to-U editing at cytidine residues. DART-seq is antibody-free, which allows for mapping m⁶A from ultra-low-input amounts of RNA. Therefore, the same group established the single-cell DART-seq (scDART-seq) (Tegowski et al., 2022) to identify RNA m⁶A sites in single cells.

2762

Chromatin accessibility and nucleosome positioning

2763

Chromatin accessibility is a widely studied characteristic of the eukaryotic genome. Open chromatin is a necessary condition for DNA to interact with other factors, such as TFs or non-coding RNAs, which play crucial roles in remodeling chromatin or initiating transcriptions. Also, the nucleosome comprises 8-unit histones and is wrapped with naked DNA to form chromatin. The movement of nucleosomes on the genome, or nucleosome positioning, affects chromatin accessibility. At the bulk level, assays for transposase-accessible chromatin (ATAC-seq) (Buenrostro et al., 2013) and Deoxyribonuclease I digestion (DNase-seq) (Song and Crawford, 2010) have been widely used to reveal that chromatin accessibility is a key component of the epigenetic landscape. The dynamics of chromatin accessibility drive cell differentiation and precise gene regulation. Profiling and analyzing chromatin accessibility at the single-cell level can help reveal the nature of cell heterogeneity and gene expression.

2764

ATAC-seq and DNase-seq have been applied to single cells, which can explore the different chromatin states and cell heterogeneity in massive cells. scATAC-seq (Buenrostro et al., 2015) combines microfluidics and Tn5 tagmentation with sequencing barcodes, while scDNase-seq utilizes fluorescence-activated cell sorting (FACS) to sort single cells and digest them with DNase I. The scDNase-seq can detect more DNase I hypersensitive sites (DHSs) with specific properties related to gene expression. However, both methods have relatively low cell throughput due to the microfluidic equipment. To improve the cell throughput in a single experiment, scATAC-seq in small volumes (μ ATAC-seq) (Mezger et al., 2018) integrates fluorescence imaging and addressable reagent deposition across a parallel nano-well array to improve the cell throughput to ~1800 cells per chip and yield higher enrichment. Another multiple index barcode method was introduced to them in single-cell profiling of chromatin accessibility by combinatorial cellular indexing (sci-ATAC-seq) (Cusanovich et al., 2015) and indexing single-cell DNase sequencing (iscDNase-seq) (Gao et al., 2021b). These approaches significantly improved the cell throughput to ~15,000 cells. Furthermore, droplet-based single-cell combinatorial indexing for ATAC-seq (dsciATAC-seq) (Lareau et al., 2019) integrated droplet-microfluidics-based method and combinatorial indexing, which makes profiling chromatin accessibility in ~500,000 single cells possible. Also, the single nucleus assay for transposase-accessible chromatin using sequencing (snATAC-seq) (Muto et al., 2021) only uses the cell nucleus for sequencing, which alleviates the mitochondrial contamination to yield higher quality cells and lower noise.

2765

Nucleosome organization and positioning are also involved in forming chromatin compaction and accessibility. Single-cell micrococcal nuclease sequencing

2800 (scMNase-seq) (Lai et al., 2018) adopts FACS sorting, lysis, and digestion by MNase to
 2801 build the library to profile genome-wide nucleosome positions. It reports cell
 2802 heterogeneity of nucleosome positioning and nucleosome spacing at DHSs.

2803

2804 *Histone modification and transcription factor binding*

2805 Different HMs indicate different chromatin states and activity of chromatin states,
 2806 which also affect TF binding and transcription. Antibody-based methods, such as
 2807 chromatin immunoprecipitation assays with sequencing (ChIP-seq) (Kim and Ren,
 2808 2006), have been widely used to profile HMs and TFs landscape on the whole genome.
 2809 Droplet-based chromatin immunoprecipitation followed by sequencing (Drop-ChIP)
 2810 (Rotem et al., 2015) and later single-cell ChIP-seq (scChIP-seq) (Grosselin et al., 2019)
 2811 first separate cells into droplets that contain lysis buffer and MNase, and then barcode
 2812 them before the immunoprecipitation step. They increase the efficiency of the
 2813 pull-down step and give low background results. To improve the read number per cell,
 2814 simultaneous indexing and tagmentation-based ChIP-seq (itChIP-seq) (Ai et al., 2019)
 2815 adopted the Tn5 transposase-based tagmentation coupled with simultaneous addition of
 2816 primers for barcoding and PCR amplification. It achieves ~9,000 reads per cell, close to
 2817 that in scATAC-seq assays. Due to the low affinity and efficiency of antibodies, all
 2818 these methods are used to profile HMs instead of TFs.

2819 Cleavage Under Targets & Release Using Nuclease (CUT&RUN) (Skene and
 2820 Henikoff, 2017) utilizes chromatin immune-cleavage on native chromatin, which is a
 2821 convenient and efficient low-input method. It has also been adapted to the following
 2822 similar techniques, including single-cell chromatin integration labeling (scChIL-seq)
 2823 (Harada et al., 2019), single-cell chromatin immune-cleavage sequencing technique
 2824 (scChIC-seq) (Ku et al., 2019), combinatorial barcoding and targeted chromatin release
 2825 (CoBATCH) (Wang et al., 2019b), antibody-guided chromatin tagmentation
 2826 sequencing (iACT-seq) (Carter et al., 2019), ultra-low-input cleavage under targets and
 2827 release using nuclease (uliCUT&RUN) (Patty and Hainer, 2021), single-cell Cleavage
 2828 Under Targets and Tagmentation (scCUT&Tag) (Bartosovic et al., 2021), and indexing
 2829 single-cell immune-cleavage sequencing (iscChIC-seq) (Ku et al., 2021). In particular,
 2830 scChIC-seq, ultiCUT&RUN, and iscChIC-seq use the protein A-micrococcal nuclease
 2831 (pA-MNase) as the cleavage enzyme, and others use the Tn5 transposase-protein A
 2832 (pA-Tn5) because of the release of MNase-cleaved fragments into the supernatant,
 2833 which is not suitable for single-cell platforms. Interestingly, Tn5-based approaches,
 2834 including CoBATCH, ultiCUT&RUN, and scCUT&Tag, profile not only the histone
 2835 modifications but also several abundant TFs, such as RNA polymerase II (POL II),
 2836 NANOG, OLIG2, and RAD21.

2837 Single-cell DNA adenine methyltransferase identification (scDamID) (Kind et al.,
 2838 2015) was applied to the detection of how the chromosomes are spatially organized
 2839 inside interphase nuclei. DNA adenine methyltransferase (Dam) methylates adenines
 2840 that are adjacent to positions where the protein of interest interacts with the DNA.
 2841 These methylated adenines are amplified by PCR and identified by NGS. Combining
 2842 single-cell DamID with messenger RNA sequencing (scDam&T-seq) (Rooijers et al.,
 2843 2019) successfully profiled the RING1B binding sites paralleling the transcriptome,

2844 providing a powerful tool to identify protein-mediated mechanisms that regulate
 2845 cell-type-specific transcriptional programs in dynamic processes and heterogeneous
 2846 tissues.

2847

2848 *3D genome structure*

2849 Chromatin is spatially and structurally organized and compartmentalized in the cell
 2850 nucleus, contributing to the effects of cis-regulatory elements (CRE) and
 2851 trans-regulatory factors. Chromosome conformation capture (3C) (Hag ège et al., 2007)
 2852 detects genomic regions located in close proximity to each other. With the continuous
 2853 development of conformation-based techniques, high-throughput sequencing-based
 2854 Hi-C has enabled genome-wide chromatin interaction detection. Similar to other
 2855 single-cell sequencing approaches, the isolation or barcoding of individual cells is a
 2856 primary task for single-cell Hi-C (scHi-C) (Nagano et al., 2013). scHi-C reduces the
 2857 scale of the traditional Hi-C protocol and sorts cells into multi-well plates for
 2858 fragmentation. Single-nucleus Hi-C (snHi-C) (Flyamer et al., 2017) amplifies the entire
 2859 genome and eliminates the biotin fill-in step. Diploid chromatin conformation capture
 2860 (Dip-C) (Tan et al., 2018) simplifies the experimental protocols with a
 2861 fragmentation-based strategy. Combinatorial indexing was introduced in Single-cell
 2862 combinatorial indexed Hi-C (sciHi-C) (Ramani et al., 2017), avoiding the need to
 2863 isolate cells. To capture long-range and higher-order interactions that are limited by
 2864 proximity ligation, single-cell split-pool recognition of interactions by tag extension
 2865 (scSPRITE) (Arrastia et al., 2022) detects both inter- and intra-chromosomal
 2866 interactions and more DNA contacts per cell.

2867 The key distinctions, limitations, and biological materials used in the original
 2868 research of the reviewed techniques are summarized in Figure 6 and Table S6. Several
 2869 challenges need to be overcome in the future. First, due to the low rate of DNA capture
 2870 and lower DNA content than RNA in a single cell, single-cell epigenome data is
 2871 currently highly sparse. Second, existing methods still have difficulty detecting the
 2872 precise binding location of TFs, particularly for TFs that are not evenly distributed over
 2873 the whole genome. Third, the elaboration of the mechanism of gene regulation from
 2874 DNA to cell states and phenotypes continues to demand the further development of
 2875 single-cell multi-omic approaches.

2876

2877 *Computational methods for single-cell epigenome data*

2878 *Reads preprocessing, quality control, and quantification*

2879 The read adaptor trimmers and mappers, which are designed for bulk tissues, can also
 2880 be used for single-cell reads. Fastp (Chen et al., 2018), and Trimmomatic (Bolger et al.,
 2881 2014) are used for removing the adapter sequence to facilitate the read mapping. For
 2882 DNA methylation data, especially data generated by bisulfite-based methods, Bismark
 2883 (Krueger and Andrews, 2011), BSMAP (Xi and Li, 2009), and Bsseeker (Chen et al.,
 2884 2010) were adopted to map the reads to the genome. Bisulfite conversion induces
 2885 largely depleted cytosines of the genome sequences, which causes multiple mapping
 2886 sequencing reads, and this situation is more serious when it comes to single-cell data.
 2887 The scBS-map (Wu et al., 2019a) was developed by remapping chimerical reads, which

is the majority of the unmapped reads, with a local alignment approach, and dramatically improving the overall mapping efficiency. For scATAC-seq or other non-converted DNA sequences, BWA (Li and Durbin, 2009), bowtie2 (Langmead et al., 2019), and minimap2 (Li, 2018) were widely used to perform the mapping. Recently, chromap has brought pseudo-alignment to DNA mapping, which significantly improves the mapping efficiency with comparable mapping rate, and has been adopted in several analysis pipelines.

For quality control (QC), FastQC is often used to control the quality at the reading level. Limitations on the number of mapped reads and mitochondrial reads per cell filter out low-quality cells. For single-cell DNA methylation, the count matrices are built from cytosine summary tables or any custom-defined features of interest. The methylation status of cytosines in CG, CH, or both genomic contexts in every feature is counted and summarized in the matrices. MethylStar (Shahryary et al., 2020) and EpiScanpy (Danese et al., 2021) both have a built-in function for quantifying the methylation reads. BPRmeth introduced generalized linear model (GLM) regression to quantify methylation profiles. For scATAC-seq, the count matrices take mapped BAM files or fragment files, like 10x Cell Ranger output, as input. There are two mainstream ways to define the features. The first solution is that the cells that pass read-level QC are merged to call peak with software that is used in bulk, like MACS2 (Zhang et al., 2008) or chromHMM (Ernst and Kellis, 2012). The peak file is regarded as the region of interest and is used to count the reads in each peak. This solution significantly reduces the feature number that accelerates the downstream analysis but may lose the information and heterogeneity of the rare cell population. Dr.seq2 (Zhao et al., 2017), MAESTRO (Wang et al., 2020), scitoools (Sinnamon et al., 2019), APEC (Li et al., 2020a), and Signac (Stuart et al., 2021) use the merged cell peak as features. Another solution is to count the reads with a segmented genome or so-called bin-based. SnapATAC (Fang et al., 2021) adopts this strategy to capture the rare population but generates numerous features that need to be carefully filtered in the downstream analysis. The count matrices for scATAC-seq are often binarized because of only double DNA strands in a cell.

2918

2919 *Imputation*

2920 As previously mentioned, the epigenome data from single cells is extremely sparse, 2921 which impacts the sensitivity and accuracy of downstream analysis for biological 2922 findings. Many methods have been developed for predicting and fulfilling the missing 2923 values as a result of bias from techniques.

2924

For DNA methylation, predicting missing methylation states and improving the incomplete CpG coverage is critical to analyzing genome-wide methylation status. DeepCpG (Angermueller et al., 2017) utilizes convolutional neural networks (CNN) to learn the associations between DNA sequence features and methylation states between neighboring CpG sites, both within a cell and across cells. MOFA (Argelaguet et al., 2018) and MOFA+ (Argelaguet et al., 2020) infer an interpretable low-dimensional data representation with principal component analysis (PCA) to impute the missing values and assays. MELISSA (Kapourani and Sanguinetti, 2019), scMET (Kapourani

et al., 2021), and Epiclomal (C et al., 2020) use Bayesian mixture models to leverage similar methylation patterns in similar cells and impute missing values.

For scATAC-seq, ChromA (Gabitto et al., 2020) also adopts a Bayesian statistical approach with hidden semi-Markov models (HSMM) to overcome the sparsity from scATAC-seq data. ScOpen integrates an unsupervised learning model based on a non-negative matrix factorization (NMF), which does not require making to make assumptions about the data distribution. AtacWorks (Lal et al., 2021) uses the ResNet (residual neural network) architecture to train a deep learning model from high-quality bulk ATAC-seq datasets and predicts improved signal tracks at the base-pair level and the accessible genomic locations with noisy scATAC-seq tracks. SCATE (Ji et al., 2020b) integrates co-activated peaks, similar cells, and publicly available bulk data to predict the signals of each peak. These imputation methods also enhance cell clustering performance.

For scHi-C, scHiCluster (Zhou et al., 2019a) considers chromosome interactions as a network and uses the random walk algorithm to propagate the smoothed interactions to tackle the sparsity of data. HiCImpute (Xie et al., 2022) considers the spatial dependencies of 2D data structure and borrows information from similar single cells and bulk data. scHiCEmbed (Liu and Wang, 2022) borrows the scHiCluster's result and uses graph auto-encoders to learn node embeddings, which enables the imputation of the chromosome contact matrices and topologically associating domains (TAD) detection. Higashi (Zhang et al., 2022e) transforms the scHi-C data into a hypergraph and imputes the scHi-C contact maps by predicting missing hyperedges within the hypergraph. Another imputation task for scHi-C is to reconstruct the 3D genome structures. Si-C (Meng et al., 2021) applies the Bayesian theory framework to reconstruct genome 3D structures from scHi-C data. SCL (Zhu and Wang, 2019) regards the 3D structure of a chromosome as beads-on-a-string and reconstructs the structure inside a 3D cubic lattice, and uses 2D Gaussian imputation to estimate the propensity for the bead-pairs without scHi-C contacts. Also, a data-driven method, SIMBA3D (Rosenthal et al., 2019) first utilizes bulk Hi-C data to aid in recovering the interactions missed in scHi-C contact maps, then infers 3D chromosome structures with a generalized Bayesian framework.

2963 *2964* Clustering

Clustering similar cells together assigns identities to cells to better find the rare cell population, understand the gene regulatory patterns in specific cell states, and alleviate the noise signals. The clustering algorithms, such as tSNE (Laurens and Hinton, 2008), UMAP (McInnes and Healy, 2018), graph abstraction (Wolf et al., 2019), Louvain clustering (Fortunato, 2009), Leiden clustering (Guo et al., 2019), and diffusion pseudotime (Haghverdi et al., 2016), which are used in single-cell transcriptomes, have been applied to single-cell epigenomes too. ALLools (Liu et al., 2021), EpiScanpy (Danese et al., 2021), Signac (Stuart et al., 2021), ArchR (Granja et al., 2021), SnapATAC (Fang et al., 2021), and other analysis pipelines integrated these algorithms as built-in functions to facilitate easy clustering of cells.

Although the clustering algorithms used in the single-cell epigenome are similar to

2976 those used in single-cell transcriptome data, single-cell epigenome data suffers from
 2977 more sparse and numerous features. To overcome sparsity, the imputation methods
 2978 mentioned in the last section can be used to improve the clustering performance by
 2979 fulfilling the missing features as well as keeping the cell heterogeneity. scABC
 2980 (Zamanighomi et al., 2018) tries to alleviate the noise from cells with low sequencing
 2981 depth by implementing a weighted version of the K-medoids clustering algorithm,
 2982 which gives a low weight to the low sequencing depth cells.

2983 Another difference in clustering algorithms for single-cell epigenome data from
 2984 single-cell transcriptome data is the reduction of features or dimensions. PCA is the
 2985 most commonly used method to reduce the dimensions of cluster features. Seurat v3
 2986 (Stuart et al., 2019) incorporates latent semantic indexing (LSI) on the scATAC-seq
 2987 feature count matrix to reduce the dimensionality. CisTopic (Bravo González-Blas et al.,
 2988 2019) uses latent Dirichlet allocation (LDA) with a collapsed Gibbs sampler to identify
 2989 the cis-regulatory topics. It also facilitates the prediction of TF binding sites and
 2990 chromatin states. PeakVI (Ashuach et al., 2022) employs a deep generative model to
 2991 learn a probabilistic low-dimensional representation. ScVAEBGM (Duan et al., 2022)
 2992 integrates a Variational Autoencoder (VAE) with a Bayesian Gaussian-mixture model
 2993 (BGM) to process scATAC-seq data. It takes advantage of the BGM to estimate cluster
 2994 numbers from data.

2995 Besides only using the information from single-cell epigenome data, borrowing
 2996 the information from sequence features, bulk datasets, and single-cell transcriptome
 2997 datasets also helps with the task of clustering. Some methods developed for multi-ome
 2998 experiments, such as MAPLE (Uzun et al., 2021), scAI (Jin et al., 2020a), LIGER
 2999 (Welch et al., 2019), scMC (Zhang and Nie, 2021), and scGCN (Song et al., 2021c),
 3000 improved the clustering performance by integration with scRNA-seq. chromVAR
 3001 (Schep et al., 2017), BROCKMAN (de Boer and Regev, 2018), scFAN (Fu et al., 2020),
 3002 and scBasset (Yuan and Kelley, 2022) consider the sequence features, including motifs
 3003 or specific k-mer, to reduce the dimension from peak level to k-mer level or TF level.
 3004 Furthermore, CellWalkR (Przytycki and Pollard, 2022) integrates scATAC-seq with
 3005 cell type labels and bulk epigenetic data to better illustrate the CREs active in specific
 3006 cell types. SCRIIP (Dong et al., 2022) incorporates many bulk ChIP-seq data sets, which
 3007 also use peak set similarity to convert the feature matrix from the peak count to TF
 3008 count. These methods not only enhance the clustering performance but also provide
 3009 biological information on which peaks or sequence features are important to specific
 3010 regulatory factors.

3011 For scHi-C data, SCL and scHiCEmbed increase the clustering performance by
 3012 alleviating the sparsity of data with imputation. Recently, scHiCStackL (Wu et al., 2022)
 3013 proposed a computational framework by constructing a two-layer stacking ensemble
 3014 model for classifying cells and outperformed other methods on the task of clustering
 3015 cell types.

3016

3017 Cell type annotation and trajectory inference

3018 Even while single-cell approaches allow for the parallel analysis of genomic data
 3019 among numerous cells, we usually need to know the cell types or differentiation stages

3020 of each cluster. Annotating cells using single-cell epigenome data typically requires
 3021 inferring the gene activity to assist in distinguishing cell types. This is in contrast to
 3022 scRNA-seq, which can identify cell states by gene markers.

3023 ArchR and MAESTRO both provide statistical models to infer the gene score at
 3024 the cluster level from scATAC-seq peaks. ArchR incorporates the exponential decay
 3025 model while accounting for the expanded gene body and gene border. MAESTRO also
 3026 uses an exponential decay model but considers the exons of each gene and removes the
 3027 effects of nearby genes. Garnett borrows the methods of calculating gene activity
 3028 scores from Cicero (Pliner et al., 2018) and applies their predefined markup language
 3029 and pre-trained classifier to scATAC-seq data. Besides using inferred gene scores as
 3030 markers to annotate cells, another way is to use well-annotated bulk data as references.
 3031 SCRAT (Ji et al., 2017) compiles a regulome database consisting of ENCODE (de
 3032 Souza, 2012) DNase-seq profiles from a wide variety of cell types to infer the likely
 3033 cell type of each cell. Moreover, MAESTRO not only uses the data from the ENCODE
 3034 project but also the data from the Cistrome Data Browser (Mei et al., 2017; Zheng et al.,
 3035 2019; Zheng et al., 2020), which has collected the most comprehensive previous public
 3036 DNase-seq and ATAC-seq datasets.

3037 The transcription of RNA takes time, therefore single-cell epigenome data is more
 3038 sensitive in capturing cell differentiation events than scRNA-seq. To infer cellular
 3039 trajectories, STREAM (Chen et al., 2019b) first uses PCA to extract the most
 3040 informative features. Modified Locally Linear Embedding (MLLE), a non-linear
 3041 dimensionality reduction technique, is then used to project cells into a low-dimensional
 3042 space before the implementation of the Elastic Principal Graph. MIRA (Lynch et al.,
 3043 2022) uses topic modeling to infer cell states and represent those states in an
 3044 interpretable latent space, allowing for the inference of cell state trees and the
 3045 identification of important regulators of branch point fate decisions. Also, many
 3046 pipeline tools, like EpiScanpy and Signac, incorporate PAGA (Wolf et al., 2019) or
 3047 Monocle (Trapnell et al., 2014) to infer the cell trajectories. However, understanding
 3048 the biological system as well as the underlying assumptions is necessary when
 3049 modeling trajectories using single-cell data. Therefore, to interpret the results of
 3050 trajectories, well-annotated clustering is often a requirement.

3051 *Differential analysis and features selection*

3052 With differential analysis, it is crucial to determine which features are related to
 3053 particular cell states. This approach connects cell states and phenotypes to genomic
 3054 regions or CREs. A recent report claimed that the Wilcoxon rank-sum test outperforms
 3055 other differential test methods in large-sample-size data because it does not require any
 3056 assumptions (Li et al., 2022i). In fact, the Wilcoxon rank-sum test is the most
 3057 commonly used test method for detecting differential expression genes in the majority
 3058 of pipeline tools.

3059 Although it is not difficult to perform the differential analysis with current tools, a
 3060 tricky thing is how to define the useful features of single-cell epigenome data.
 3061 Bin-based methods and peak-based methods are adopted for scATAC-seq. scMET
 3062 aggregates the input data within regions, such as promoter regions or enhancers. These

3064 genome features rely on the aggregation of individual regions. Recently, a deep
 3065 generative model PeakVI infers a representation for each cell in high-dimensional,
 3066 which enables statistically robust inference of single-region-level differential
 3067 accessibility and cell state annotation.

3068

3069 *Gene regulation inference*

3070 Inferring TF activity using single-cell epigenomics data is an intriguing potential
 3071 application that provides clues on how epigenetics influences gene expression and cell
 3072 phenotypes. ChromVAR, scFAN, scBasset, TRIPOD (Jiang et al., 2022b), and SCRIP
 3073 all support inferring TF activity at the single-cell level from scATAC-seq data.
 3074 ChromVAR infers TF activity by estimating the gain or loss of accessibility within
 3075 peaks sharing the same TF motifs. scFAN pre-trains deep learning-based models on
 3076 genome-wide bulk ATAC-seq, DNA sequence, and ChIP-seq data and applies the
 3077 model to single-cell ATAC-seq to predict TF binding in individual cells. scBasset
 3078 introduces CNNs to leverage the DNA sequence information underlying scATAC-seq
 3079 peaks to achieve TF activity inference. TRIPOD combines scRNA-seq, scATAC-seq,
 3080 and DNA sequence features to infer the TF activity related to gene expression
 3081 associations, accounting for literature-based knowledge. However, the DNA sequence
 3082 features, such as motifs, lose the cell-type-specific information of TFs and cannot
 3083 distinguish between TFs with similar motifs, such as the GATA family. Recently,
 3084 SCRIP incorporated thousands of bulk-level ChIP-seq datasets and scATAC-seq to
 3085 infer the TF activity based on the peak set similarity, which successfully distinguished
 3086 the similar motif TF activity at the single-cell level.

3087 Although scATAC-seq identifies the open chromatin regions as CREs, how the
 3088 CREs link distal regulatory elements with their target genes is also a key question in
 3089 gene regulation. Cicero samples and aggregates similar cells to quantify correlations
 3090 between putative CREs and links CREs to target genes based on the correlation using a
 3091 graphical lasso model. To alleviate uncorrelated technology noise and false positive
 3092 results in Cicero, JRIM (Dong and Zhang, 2021) uses the group lasso penalty to find
 3093 similar patterns of sparsity across all the regulatory networks to reconstruct the
 3094 cis-regulatory interaction networks. To accurately identify the loci of key CREs of
 3095 different cell types, scEpiLock (Gong et al., 2022) adopts a CNN model to detect the
 3096 chromatin accessibility regions and refine the peak boundary using gradient-weighted
 3097 class activation mapping (Grad-CAM). Similarly, DIRECT-NET (Zhang et al., 2022c)
 3098 adopts eXtreme Gradient Boosting (XGBoost) to identify functional CREs and infer
 3099 the TF binding sites with known motif patterns. The aforementioned methods
 3100 successfully link the CREs to target genes, DeepTFni (Hao et al., 2022) implements a
 3101 graph neural network (GNN) with a variational graph auto-encoder (VGAE) to infer TF
 3102 regulatory networks, which can show the relationship between TFs and TFs. SMGR
 3103 (Song et al., 2022b) takes both scRNA-seq and scATAC-seq as input and utilizes a
 3104 generalized linear regression model to identify the latent representation of consistently
 3105 expressed genes and peaks, as well as identify co-regulatory mechanisms.

3106 ScHi-C allows for exploring gene regulation patterns in a 3D manner at the
 3107 single-cell level. Topologically associating domains (TAD) segment the genome based

3108 on the 3D genome structure. There are more DNA-DNA interactions within TADs than
 3109 between one TAD and other TADs. deTOKI (Li et al., 2021c) can predict TAD-like
 3110 domain structures at the single-cell level with NMF from sparse scHi-C data.
 3111 Chromatin loops are smaller structures that link CREs to target genes physically.
 3112 SnapHiC (Yu et al., 2021) and SnapHiC2 (Li et al., 2022h) enable identifying
 3113 chromatin loops at 10 kb resolution with a random walk with restart (RWR) algorithm
 3114 from scHi-C data.

3115

3116 *Multi-function analysis pipelines*

3117 The selection and organization of these tools to effectively extract the underlying
 3118 information from data have become a challenge with the development of numerous
 3119 computational approaches for single-cell epigenomic data. For example, Chen et al.
 3120 benchmarked 10 computational methods that were developed for scATAC-seq and
 3121 concluded that different methods have their advantages and limitations (Chen et al.,
 3122 2019c). Multi-function pipelines provide one-shot solutions with parameters based on
 3123 best practices, freeing biologists from menial coding and parameter tuning so that they
 3124 can focus on the biological results.

3125 Dr.seq2, SCRAT, Scasat, Destin, scitoools, scATAC-pro, EpiScanpy, Signac, and
 3126 SnapATAC are designed especially for single-cell chromatin accessibility or
 3127 methylome accessibility. scHiCTools is a pipeline that is designed for scHi-C data.
 3128 They include the functions of basic qualification, filtering low-quality cells or features,
 3129 motif analysis, clustering, differential analysis, and visualization. Since many
 3130 techniques have been developed for parallelly profiling the transcriptome and
 3131 epigenome, many computational methods and pipelines have been developed for
 3132 integration. Seurat v3, APEC, MAESTRO, scAI, ArchR, and ALLCools provide the
 3133 functions that are mentioned above as well as functions for integration of the
 3134 epigenome data and transcriptome to better interpret the gene regulation mechanism.

3135 Besides these computational methods, g-chromVAR (Ulirsch et al., 2019) uses
 3136 fine-mapped variant posterior probabilities and quantitative measurements of
 3137 regulatory activity to measure the enrichment of regulatory variants in each cell state.
 3138 Methylscaper (Knight et al., 2021) is specifically developed for visualizations of
 3139 single-cell DNA methylation and chromatin accessibility patterns. Several integration
 3140 methods have been developed to analyze scRNA-seq and single-cell epigenome data
 3141 together. These include MATCHER (Welch et al., 2017), coupled NMF (Duren et al.,
 3142 2018), coupleCoC (Zeng et al., 2021), coupleCoC+ (Zeng and Lin, 2021), scAMACE
 3143 (Wangwu et al., 2021), epiConv (Lin and Zhang, 2022), scMVP (Li et al., 2022b),
 3144 scREG (Duren et al., 2022), and MIRA. These computational methods for integration
 3145 provide a more thorough and multifaceted perspective in which to understand the gene
 3146 regulatory process. Table S7 lists the programming language, key features, limitations,
 3147 and benchmark dataset that were applied in the original analysis of the reviewed
 3148 computational approaches (Figure 7).

3149

3150 *Applications of single-cell epigenomes*

3151 Single-cell technologies provide unprecedented opportunities to investigate a variety of

3152 biological processes and gene regulation patterns. Applying these single-cell
 3153 technologies to various biological systems sheds light on discovering the cell
 3154 differentiation events and mechanisms of disease occurrence at the single-cell level.
 3155 These single-cell epigenome sequencing methods have been adopted in many fields.
 3156 Here, we reviewed their applications in early embryonic development, cancer, and
 3157 neurobiology.

3158

3159 *Early embryonic development*

3160 During gamete development and the early stages of embryogenesis, cells undergo
 3161 significant and drastic alterations and reprogramming in the epigenome, which causes
 3162 cell differentiation and diverse phenotypes of cells. Therefore, embryonic stem cells are
 3163 widely used as material in the development of single-cell epigenomic sequencing
 3164 techniques.

3165 Zhu et al. applied scWGBS to human preimplantation embryos (Zhu et al., 2018a).
 3166 They discovered three waves of global demethylation in mouse preimplantation
 3167 embryos, indicating that the dynamic balance between global demethylation and drastic
 3168 remethylation occurs during preimplantation development. Later, the same group, Li et
 3169 al. (Li et al., 2018a) applied scCOOL-seq to six stages of human preimplantation
 3170 development and discovered that the pluripotency master TF binding regions and
 3171 proximal and distal nucleosome-depleted regions were primarily enriched in the
 3172 genomic regions showing the largest changes in chromatin accessibility. Additionally,
 3173 they discovered that, compared to mice, human zygotes had reduced access to the
 3174 maternal genome's chromatin in oocytes and had a delayed balance between parental
 3175 alleles until the 4-cell stage, which indicated the species-specific features of chromatin
 3176 accessibility. Argelaguet et al. (Argelaguet et al., 2019) performed scNMT-seq on the
 3177 stages of mouse gastrulation. They found that cells committed to mesoderm and
 3178 endoderm undergo widespread coordinated epigenetic rearrangements at enhancer
 3179 marks, driven by ten-eleven translocation (TET)-mediated demethylation and a
 3180 concomitant increase in chromatin accessibility. In addition, they found that while in
 3181 the early epiblast, the methylation and accessibility landscape of ectodermal cells had
 3182 already been established.

3183 These studies shed light on how the epigenome influences cellular differentiation
 3184 and lineage commitment. In the future, investigations into cell populations using
 3185 single-cell multi-omics techniques give us the chance to understand the process of
 3186 orchestrated epigenomic reprogramming, which has the potential to change our
 3187 understanding of cell fate decisions and benefit the field of stem cell biology.

3188

3189 *Tumor immunology*

3190 Malignant and non-malignant cells coexist in a tumor, which is a highly heterogeneous
 3191 structure. Both types of cells play critical roles in the development of cancer. Methods
 3192 for single-cell epigenome sequencing are being developed to help distinguish the
 3193 non-genetic factors that contribute to the course of cancer from the complexity of
 3194 tumors.

3195 Satpathy et al. (Satpathy et al., 2019) applied scATAC-seq on primary tumor

3196 biopsies from basal cell carcinoma (BCC) patients receiving PD-1 blockade treatment.
 3197 They investigated chromatin regulators of therapy-responsive T cell subsets and
 3198 observed a common regulatory pathway that controls the development of CD4⁺ T
 3199 follicular helper cells and intratumoral CD8⁺ T cell exhaustion. Not only are immune
 3200 cells investigated by single-cell epigenome sequencing, but also malignant cells show
 3201 heterogeneities in TME. Meir et al. (Meir et al., 2020) employed scRNA-seq and
 3202 methylome analysis to show that various cancer cell types had clonally stable
 3203 epigenetic memory. Additionally, they discovered DNA methylation landscapes reflect
 3204 a separate clock-like methylation loss mechanism while correlating with
 3205 epithelial-to-mesenchymal transcriptional (EMT) identities that are identified by
 3206 transcriptome analysis in clonal colon cancer cell populations. Wu et al. (Wu et al.,
 3207 2021c) employed scCUT&Tag to characterize H3K27me3 before and after therapy in a
 3208 patient with a brain tumor. They profiled a brain tumor H3K27me3 in the primary
 3209 sample and after the treatment and discovered various cell types in the TME and
 3210 heterogeneity in the polycomb group activity.

3211 Epigenetic mechanisms are critical for the interactions between tumor cells and
 3212 immune cells. Understanding the fundamental processes of epigenetic modifications in
 3213 immune and tumor cells paves the way for the creation of drugs and immunotherapy
 3214 techniques.

3215 *Neurobiology*

3216 Understanding both the normal functions of the brain and the mechanisms of
 3217 dysfunction and disease requires a better understanding of cellular composition. Lake
 3218 et al. (Lake et al., 2018) detected the transposon hypersensitive sites in the human adult
 3219 brain at the single-cell level. They identified the cell subpopulations in the human adult
 3220 cortex and cerebellar hemisphere and used epigenomic data to link genetic risk variants
 3221 with cell-type-specific cCREs. In a cohort of cognitively healthy people, Corces et al.
 3222 (Corces et al., 2020) examined the single-cell chromatin accessibility landscapes and
 3223 three-dimensional chromatin interactions of various adult brain regions. They created a
 3224 machine-learning classifier to include this multi-omic framework and predicted several
 3225 functional SNPs for Parkinson's and Alzheimer's disease. Yang et al. (Yang et al.,
 3226 2023a) profiled single nucleus-accessible chromatin landscape of the pig hippocampus
 3227 at different developmental stages and revealed notable enrichment of transposable
 3228 elements in cell type-specific accessible chromatin regions. This study helps deepen
 3229 our understanding of human neurodegenerative diseases. Future research on the
 3230 single-cell level will be fascinating in examining dynamic regulations of the epigenome,
 3231 specifically alterations to the genome during learning and memory that are reliant on
 3232 neuronal activity.

3233 *Summary*

3234 In this chapter, we summarized the techniques, computational methods, and
 3235 applications for single-cell epigenome sequencing. The recent availability of single-cell
 3236 sequencing technology has expanded the scope of study into biological processes and
 3237 diseases. These approaches have previously shown their effectiveness in illuminating

3240 the parts of complex tissues and revealing novel insights, despite some limitations.
 3241 Future sequencing technologies with higher coverage and sensitivity, as well as
 3242 dedicated, advanced, and well-developed computational methodologies, promise to
 3243 usher in a new era of understanding biology and pave the way for the treatment of
 3244 diseases.

3245

3246 **Chapter 4 Single-cell proteomics technology based on mass spectrometry**

3247

3248 Although proteome performed in bulk samples has matured and been widespreadly
 3249 applied in scientific research and clinical medicine, numerous new challenges have
 3250 arisen when the initial sample size is slashed to hundreds of cells or even one cell.
 3251 Since the total protein in one somatic cell is only about 100~200 pg (Wiśniewski et al.,
 3252 2014), any loss can have an immeasurable impact, and also put more stringent
 3253 requirements on sensitivity and accuracy of detection methods.

3254 Based on the detection principles, today, single-cell proteomic (SCP) methods
 3255 can be classified into the following categories: (1) antibody-based assays, such as
 3256 cytometry by time of flight (CyTOF) (Bandura et al., 2009; Bendall et al., 2011),
 3257 single-cell western blotting (Hughes et al., 2014b), microengraving (Love et al., 2006;
 3258 Han et al., 2010; Schubert et al., 2016), (2) PCR-sequencing based assays, such as
 3259 proseek multiplex (Assarsson et al., 2014), CITE-seq (Stoeckius et al., 2017),
 3260 REAP-seq (Peterson et al., 2017), (3) mass spectrometry (MS) based assays, such as
 3261 nanoPOTS (nanodroplet processing in one pot for trace samples) (Zhu et al., 2018d;
 3262 Zhu et al., 2018e), SCoPE2 (Specht et al., 2021). The antibody-based and
 3263 sequencing-based assays were developed earlier and have made unignorable
 3264 contributions, which have been well reviewed before (Levy and Slavov, 2018; Labib
 3265 and Kelley, 2020; Liu et al., 2020a; Xie and Ding, 2022). However, the former was
 3266 limited by the specificity and the availability of antibodies, while the latter did not
 3267 directly detect proteins, and both assays were limited in the number of proteins that
 3268 could be analyzed at a time. Mass spectrometry has been the mainstream analysis tool
 3269 in bulk-size proteomics due to its advantages of both high accuracy and high
 3270 throughput. More importantly, unlike the hypothesis-oriented antibody-based methods,
 3271 MS-based analysis method is discovery-oriented which can play a unique role in
 3272 biological research. With the progress of mass spectrometry and the innovation of
 3273 preparation process in recent years, single-cell proteomics based on MS has shown a
 3274 blowout.

3275 Here we focus on the state-of-the-art MS-based single-cell proteomic tools over
 3276 the last 5 years, discuss the outstanding innovative points from cell isolation to sample
 3277 preparation and MS detection, and prospect the future development directions.

3278

3279 *MS-based SCP Workflow*

3280 The workflow of conventional bulk-size proteomics has been well established, but
 3281 when dealing with the single-cell, each simple step needs to be well optimized to
 3282 recover as much information as possible from the trace protein. Although the ways of
 3283 implementation vary, the workflows of MS-based SCP include three main steps:

3284 single-cell isolation, sample preparation, and MS analysis.

3285

3286 *Single-cell Isolation*

3287 Isolation of single-cell is a unique requirement in SCP workflows, its accuracy and
3288 efficiency lay the foundation for the entire SCP workflow. How to precisely pick the
3289 interesting cell types from a complicated cell mixture, and how to maximize cell
3290 activity and minimize the impact on the molecular level are the great challenges in
3291 this step. Among the existing SCP tools, isolation of single-cell was mostly achieved
3292 through manual cell picking, fluorescence-activated cell sorting (FACS), microfluidic,
3293 lab-on-a-chip devices (Gross et al., 2015), or Laser capture microdissection (LCM)
3294 (Mund et al., 2022a).

3295

3296 *Sample preparation*

3297 Cell lysis and protein digestion are the essential preparation steps for both bulk-size
3298 and single-cell proteomic workflows. Since the protein amount contained in one cell
3299 is extremely small, any step that may be negligible for bulk-size proteomics can be
3300 fatal to the SCP. Even the nonspecific adsorption during sample transfer steps can
3301 cause massive protein and peptide loss (Wu et al., 2019b; Sun and Kumar, 2022).
3302 Across the existing SCP tools, the improvement strategies can be summarized into (1)
3303 simplifying the preparation steps, (2) minimizing the sample volume, and (3)
3304 automating the preparation.

3305 In bulk-size proteomics, harsh chemical environments such as sodium dodecyl
3306 sulfonate (SDS) and urea were required for millions of cell lysis, which need complex
3307 buffer exchange steps followed to adapt the enzymatic processes and liquid
3308 chromatography-mass spectrometry (LC-MS) analysis. To avoid loss during these
3309 steps, most SCP tools chose MS compatible lysis reagents such as DDM
3310 (n-Dodecyl-β-D-Maltopyranoside), RapiGest, trifluoroethanol (TFE) (Li et al., 2018b;
3311 Zhu et al., 2018c; Zhu et al., 2018d; Williams et al., 2020; Schoof et al., 2021;
3312 Czortecka et al., 2022a; Wang et al., 2022d), and supplemented with heating or
3313 sonication to promote cell lysis. Slavov's group developed a sample preparation
3314 method called mPOP which can lyse mammalian cells in pure water by a freeze-heat
3315 cycle (-80 °C to 90 °C), completely avoiding the introduction of redundant chemical
3316 reagents (Specht et al., 2021). Moreover, almost all SCP tools took the one-pot
3317 preparation to avoid loss during sample transfer, or even completing overall
3318 preparation in highly integrated microfluidic chips or capillary, such as iPAD-1 (Shao
3319 et al., 2018), iProChip (Gebreyesus et al., 2022).

3320 Decreasing sample volume is another solution to the loss of nonspecific
3321 adsorption, which can increase sample concentration meanwhile. Compared with
3322 bulk-size proteomics, the protein amount in SCP has a thousand-fold decrease, but the
3323 sample volume has not decreased to the same extent, leading to a great reduction of
3324 protein concentration. When the concentration is low enough, the tiny amount of
3325 protein in the large volume of solution can cause “swimming pool effect” that
3326 dramatically reduces the reaction efficiency of enzyme or label reagent with proteins
3327 or peptides. Increasing the concentration of enzyme or label reagent can partially

alleviate this effect, but an excess of the enzyme or label reagent relative to the protein or peptide level will inevitably cause signal interference in downstream analysis and increase costs exponentially. Therefore, minimization of sample volume can greatly solve the interference caused by insufficient sample concentration and significantly enhance SCP performance. Most recent SCP tools did not exceed the initial volume of 1 μ l in the sample preparation step. With the help of capillary-based sampling method or picoliter-level liquid dispensing technology, the initial volume can be controlled to as low as 2 nl or 8 nl, and the overall reaction system volume is not greater than 50 nl during the entire preparation process (Leduc et al., 2022). When the sample volume is only a few microliters or even tens of nanoliters, how to avoid liquid evaporation during the preparation process becomes a new problem. Precise control of temperature and humidity throughout the reaction environment is necessary. Oil or water-insoluble organic solvent is also used to encapsulate protein droplets to prevent volatilization. For example, OAD (nanoliter-scale Oil-Air-Droplet) used fluorinated oil FC40 to cover the cell sample (Li et al., 2018b), proteoCHIP chose hexadecane which solidified at preparation temperature (Ctortecka et al., 2022a), and WinO designed the entire preparation process in the droplet coated with ethyl acetate (Masuda et al., 2022).

As the sample size decreases, the perturbations caused by human manipulation of the entire SCP workflow are further amplified. To improve throughput and micromanipulation accuracy, some laboratories have developed specific robots to complete the integrated preparation process, such as nanoPOTS dispensing robot (Zhu et al., 2018d) and SODA (the Sequential Operation Droplet Array) system used in OAD (Zhu et al., 2013; Li et al., 2018b), as well as other SCP tools employing commercial robotic workstations to cover part or whole preparation steps. The automation of SCP sample preparation has been well reviewed by Alexović et al. (Alexović et al., 2021) Automation is an inevitable trend to further improve throughput and reproducibility so that single-cell proteomes can be truly applied to large-scale investigations or clinical studies.

MS analysis

The complexity of proteins has always been one of the biggest barriers in proteomic research. When MS-based proteomics is applied to the single-cell field, the extremely tiny sample size presents new challenges. The types of protein can reach over 10,000 with different expression levels and different physicochemical properties in one cell (Zhang et al., 2013). How to minimize the loss during MS analysis process without reducing the resolution of complex samples, and how to improve the sensitivity and accuracy of detection meanwhile are huge problems that need to be considered from three aspects: analysis strategy, injection method, and chromatography-mass spectrometry performance.

Analysis strategy

Isobaric label-based quantification, such as tandem mass tags (TMT), has been one of the most popular protein quantification methods. TMT label-based bulk-size

proteomics has been shown to result in a 15%~20% increase in proteins identified with higher quantitative accuracy (Muntel et al., 2019). Slavov's group developed the SCoPE-MS (Single-Cell ProtEomics by Mass Spectrometry) which pioneered the application of TMT to single-cell proteomics (Budnik et al., 2018). They introduced the idea of "carrier channel" which consisted of about two hundred cells to share most of the loss from single-cell channels caused by non-specific adsorption. Meanwhile, the carrier channel provided the most signal for MS analysis, reducing the required sensitivity 10- to 100-fold. Several SCP tools have been developed subsequently based on isobaric labeling, including SCoPE2, WinO, and so on (Specht et al., 2021; Masuda et al., 2022). The improvement in MS analysis throughput is another advantage of isobaric labeling SCP tools that cannot be ignored. Labeled with the TMTpro18-plex, one injection can analyze more than 14 single cells (Leduc et al., 2022). The appropriate number of cells used in carrier channel, however, is still debatable. Cheung et al. revealed that high levels of carrier channels may adversely affect quantitative accuracy (Cheung et al., 2021). With the fast development of mass spectrometer, higher detection sensitivity helped reduce the number of cells in the carrier channel. Using 25 cells as carrier channels or eliminating carrier channels, proteoCHIP identified an average of 1,812 or 1,477 proteins from one mammalian somatic cell respectively (Ctortecka et al., 2022a).

The label-free SCP tools that analyze one cell at a time are the equally important development direction in the field of single-cell proteomics. Data-independent acquisition (DIA) is becoming mainstream in bulk-sized proteomics because of its accurate quantification with low missing values and high analytical depth. With the generation of a project-specific library, DIA mode can help label-free proteomics achieve a much higher analytical depth. DIA mode has begun to be applied to label-free SCP tools recently (Brunner et al., 2022; Gebreyesus et al., 2022; Wang et al., 2022d). Compared with the most data-dependent acquisition (DDA)-based label-free SCP work which identified about 1,000 proteins, DIA mode helped identified protein numbers rise up to more than 2,000 in one single cell. Using the same SCP workflow to detect the same type of cells, DIA mode can increase the protein identification number of one single cell by up to 188% compared with is not compatible with the common perception, some groups have already tried to develop the multiplexDIA method such as DIA-TMT and plexDIA to improve data integrity and reliability without reducing throughput (Ctortecka et al., 2022b; Derkx et al., 2022).

3407

3408 *Injection method*

3409 Despite the advantages of minimizing sample volume, the nanoliter-level sample
 3410 droplet is not compatible with most commercial LC autosamplers. To solve this
 3411 problem, nanoPOTS group developed the complicated manual loading procedures that
 3412 aspirated nanodroplet samples to a section of capillary, then eluted the sample onto a
 3413 solid-phase extraction (SPE) column, and finally inserted the SPE column with an
 3414 analytical column for gradient separation and MS detection (Zhu et al., 2018d). These
 3415 procedures are not only complex and time-consuming but also highly dependent on

the proficiency of the operators. As an improvement, a nanoliter-scale autosampler integrating nanoPOTS-based sample preparation with automated LC-MS platforms was developed, which enhanced the analysis throughput based on label-free nanoPOTS from 6 cells to 24 cells one day (Specht et al., 2018). Integration of sample preparation with LC-MS analysis has been an important development trend of SCP tools for its robustness and high-throughput. Some integrated tools were based on the self-development autosamplers such as autoPOTS (Specht et al., 2018; Woo et al., 2021) and self-aligning monolithic (SAM) devices (Li et al., 2018b; Wang et al., 2022d), while others were developed based on a high-integrated microfluidic chip or device such as proteoCHIP (Ctortecka et al., 2022a), and iPAD-1 (Shao et al., 2018). Although most of the integrated SCP methods require customized equipment which limits their accessibility, from the perspective of minimizing the loss during sample loading and optimizing the detection effect, the integration of sample preparation and LC-MS analysis is still an inevitable road.

3430

3431 *Chromatography-mass spectrometry performance*

3432 The overall sensitivity of the chromatography-mass spectrometry system is crucial for
3433 the analysis of extremely tiny amounts of peptide samples. Decreasing the
3434 chromatographic flow rate and narrowing separation columns' inner diameter are
3435 widely used to enhance the separation performance and ionization efficiency. Most
3436 bulk-size proteomics conventionally uses the 75 μm i.d. reversed-phase LC columns
3437 which operate at 300 nL/min. Zhu et al. have demonstrated that using 30 μm i.d.
3438 columns operating at 50 nL/min can remarkably improve the proteome coverage and
3439 have applied to most nanoPOTS-relative work (Specht et al., 2018; Zhu et al., 2018f;
3440 Woo et al., 2021). Although narrower columns and lower flow rate were also tried, the
3441 challenges in column package and longer chromatographic gradients limited routine
3442 use. A variety of prospective LC technologies also have been explored to improve
3443 separation efficiency and have been applied in low-input proteomics, including
3444 capillary electrophoresis (Lombard-Banek et al., 2016; Lombard-Banek et al., 2019),
3445 porous layer open tubular (PLOT) columns (Li et al., 2015), monolithic capillary
3446 columns (Greguš et al., 2020), micropillar array columns (μPAC) (Stadlmann et al.,
3447 2019) and so on. It is worth looking forward to their applications in the SCP field with
3448 advanced mass spectrometry.

3449 Mass spectrometry has undergone substantial development in the past decades,
3450 reflected in the improvement of data acquisition speed, detection limit, resolution, and
3451 accuracy. Orbitrap series mass spectrometers are the most commonly used in both
3452 bulk-size and single-cell proteomics because of their outstanding performance in both
3453 resolution and accuracy. With updates to Orbitrap platforms, the protein information
3454 available from single cells has increased significantly. For example, when analyzing 2
3455 ng peptide sample, nearly 3-fold unique peptides can be identified by an Orbitrap
3456 Fusion Lumos compared with an LTQ Orbitrap XL mass spectrometer. Further,
3457 comparing the Orbitrap Fusion Lumos with a newer Orbitrap Eclipse mass
3458 spectrometer, the protein coverage from one single cell increased by about 20% (Kelly,
3459 2020). Another notable breakthrough in mass spectrometry is the introduction of ion

mobility which added a new separation dimension and resulted in the transition from 3D-Proteomics (retention time, m/z, and ion intensity) into 4D-Proteomics. TimsTOF series mass spectrometers are representative and have been applied in several recent SCP tools such as PiSPA (Wang et al., 2022d), UE-SCP (Gu et al., 2022b), and T-SCP (Brunner et al., 2022). Combined with parallel accumulation–serial fragmentation (PASEF), timsTOF can achieve almost 100% ion utilization and more than 10-fold increase in sensitivity (Meier et al., 2018). FAIMS Pro™ interface is another popular technique to combine ion mobility with mass spectrometry and can be used in conjunction with Orbitrap series mass spectrometers (Shvartsburg et al., 2006). Applied in the SCP field, Field Asymmetric Ion Mobility Spectrometry (FAIMS) has been shown to increase protein coverage by 2.3-fold in a single HeLa cell (Cong et al., 2020).

3472

3473 *State-of-the-art SCP tools*

Recently a variety of SCP tools have sprung up. With the comprehensive advance in cell isolation, sample preparation, and MS analysis mentioned above, the number of proteins identified from one cell has jumped from about 100 to more than 3,000 nowadays. The mainstream bulk-size MS-based proteomics performs complex sample preparation and off-line sample loading steps separately. The single-cell proteomics, however, developed a series of integrated tools to reduce the loss of tracing peptides during pretreatment and sample loading. As a double-edged sword, integrated tools usually require specially customized equipment which limits their promotion and application among other laboratories. Many unintegrated and easy-to-use tools have been developed at the same time (Table 2).

3484

3485 *Integrated tools*

Chen et al. established an integrated proteome analysis device called iPAD-100 in 2015 which completed the whole progress from cell preparation to injection sample into the LC-MS system (Chen et al., 2015b). iPAD-100 can accomplish cell lysis and protein digestion in a fused-silica capillary simultaneously in only 1 h and robustly identify 635 proteins from 100 living DLD-1 cells. As an updated version, iPAD-1 chose the 22 μm o.d. capillary for single-cell picking and sample preparation, reduced the reactor volume to 2 nL, and compressed the preparation time to 30 min (Shao et al., 2018) (Figure 8A). With further optimized ultrasensitive nano-LC-MS/MS system, a maximum of 328 proteins were identified from one Hela cell.

Li et al. designed a droplet-based microfluidics chip called OAD chip which is composed of 4-layer cube structure (Li et al., 2018b) (Figure 8B). With the isolation by oil in the isolation layer and oil layer, about 100 nl sample droplet can be encapsulated in droplet layer to avoid evaporation and contamination. The entire preparation process took place in the droplet layer which was manipulated through a 3D printing fabricated SAM device with cylinder geometry. The enzymolytic peptide sample was then directly loaded into the nanoliter-level separation column in a pressured manner. With this approach, 51 and 355 proteins were identified in one Hela cell and one mouse oocyte respectively. Recently, Dang et al. applied OAD to

3504 human pre-implantation embryos and achieved a median of 3,736 protein
 3505 identification from single 2-cell stage human embryos (Dang et al., 2023).

3506 NanoPOTS is another microfluidics chip based nanodroplet processing platform
 3507 that was developed by Zhu et al. (Zhu et al., 2018d). NanoPOTS chip is composed of
 3508 a nanowell-patterned glass slide, a glass spacer, and a membrane-coated glass slide
 3509 (Figure 8C). The surface area of each nanowell was only 0.8 mm², which greatly
 3510 reduced the nonspecific adsorption loss on the reaction vessel surface. Interfacing
 3511 with FACS made nanoPOTS become an excellent and robust SCP tool that can
 3512 identify 670 proteins from one HeLa cell when employed the MaxQuant match
 3513 between run (MBR) algorithm (Zhu et al., 2018c). Although the sample preparation
 3514 was accomplished in the highly integrated chip, the early nanoPOTS still needed a
 3515 complicated system to load sample to LC-MS manually. An autosampler was
 3516 developed to solve this problem soon afterward and improve the throughput from 6
 3517 cells to 24 cells per day in label-free experiments (Specht et al., 2018). TMT label was
 3518 also introduced to nanoPOTS and an improved boosting to amplify signal with
 3519 isobaric labeling (iBASIL) strategy was put forward (Tsai et al., 2020). 1,424 proteins
 3520 could be identified from a single cell by using TMT10plex label and a boost channel
 3521 containing 10ng peptides. The throughput was further increased to 77 cells per day.
 3522 Recently, nested nanoPOTS (N2) chip derived from classical nanoPOTS was
 3523 developed (Woo et al., 2021). It was designed with tighter nanowell array and
 3524 increased the cell number that can be analyzed in one chip to 243 (Figure 8D). By
 3525 further reducing the nanowell volume to about 30 nl, the protein recovery was
 3526 increased by 230%.

3527 ProteoCHIP is also a highly integrated SCP tool that was designed with two parts:
 3528 (1) a nanowell layer that included 12 fields each containing 16 nanowells, (2) a funnel
 3529 layer that can pool samples from the same TMT set via centrifugation and online
 3530 connect to LC autosampler (Figure 8E) (Ctortecka et al., 2022a). proteoCHIP
 3531 eliminated all manual sample handling steps and resulted in a high-throughput and
 3532 high sensitivity analysis, which can identify an average of 1,812 or 1,477 proteins
 3533 from one mammalian somatic cell using 25 cells as carrier channels or eliminating
 3534 carrier channels respectively. This was the first attempt to eliminate carrier channels
 3535 in TMT label-based SCP and has achieved remarkable performance.

3536 Unlike most microfluidic chips used in SCP were open, Gebreyesus et al.
 3537 designed a confined, highly integrated microfluidic chip called iProChip (Gebreyesus
 3538 et al., 2022) (Figure 8F). This chip was composed of 9 units and each of them
 3539 contained a cell capture, imaging and lysis chamber, a protein reduction, alkylation
 3540 and digestion vessel, and a peptide desalting column. Size-based single-cell capture
 3541 was achieved by the wedge-shaped twin pillar arrays and the following preparation
 3542 process can be accomplished online in the chip. With the optimized DIA-MS analysis,
 3543 1,160 proteins were identified from one PC-9 cell. SciProChip was derived from
 3544 iProChip and dedicated to 20-plex processing of single cells. It showed an
 3545 improvement in cell usage efficiency of ~40% and in protein coverage of 1.53-fold.
 3546 From one PC-9 cell, SciProChip-DIA can identify about 1,500 proteins.

3547 Recently, Wang et al. developed the pick-up single-cell proteomic analysis

(PiSPA) workflow which accomplished single-cell sorting, multi-step preparation and injection of peptides to the LC column integrally by the automated pick-up operation system based on capillary probes (Wang et al., 2022d). In order to avoid losses of sample transfer, this workflow directly dispensed single cells into a commercial insert tube, using the conical bottom tip of the insert tubes as the nanoliter microreactors for sample pretreatment of single cells (Figure 8G). These insert tubes coupled with sample vials were subsequently used as sample tubes for the autosampler of the LC system to perform the sample injection. With the help of the high sensitivity LC-timsTOF system and DIA mode, 2467 proteins can be identified from one A549 cell. PiSPA achieved the largest protein number identified from one single mammalian somatic cell in label-free SCP up to now.

3559

3560 *Easy-to-use Tools*

3561 SCoPE pioneered the application of TMT labeling strategies in single-cell proteomics
3562 and introduced a “carrier channel” containing hundreds of cells to share the most
3563 non-specific adsorption loss as well as to provide most signal for MS analysis
3564 (Budnik et al., 2018). Single cells were picked manually into reaction tubes
3565 containing H₂O and then mechanically lysed by adaptive focused acoustics (AFA).
3566 Any detergent that would interfere with the LC-MS system was not introduced. With
3567 TMT10plex labelling, 767 proteins were identified from single U-937 or Jurkat cell
3568 using 200 cells as carrier. To overcome the weaknesses of expensive AFA equipment
3569 and low-throughput manual cell isolation, a second-generation tool, SCoPE2, was
3570 developed (Specht et al., 2021). Single cells were sorted by FACS into the
3571 commercial 384-well plates containing 1 µl H₂O and then lysed through
3572 freeze-thawing (Figure 9A). Compared with SCoPE, SCoPE2 decreased lysis
3573 volumes by 10-fold, reduced the cost of consumables and equipment by over 100-fold,
3574 increased the throughput of sample preparation by over 100-fold, and increased the
3575 identified proteins from one cell up to 1,000. This was the first SCP tool that could
3576 identify more than 1,000 proteins in one cell without any customized equipment or
3577 expensive instrument. Recently, the same group developed a new preparation method
3578 called nPOP which employed cellenONE for cell sorting and liquid operation (Leduc
3579 et al., 2022). nPOP enabled the simultaneous and automated preparation of over 2,000
3580 single cells in droplets on a special fluorocarbon coated glass slide surface. Although
3581 automatic precision operations and high throughput attenuated the batch effect, nPOP
3582 became less accessible. Schoof et al. developed a similar SCP tool inspired by
3583 SCoPE2 (Schoof et al., 2021). They replaced water with TFE as the lysis reagent,
3584 which has been shown to produce more protein and especially peptide identifications.
3585 It is worth mentioning that they developed a computational workflow, SCeptre (Single
3586 Cell proteomics readout of expression), for the analysis of SCP MS data. SCeptre was
3587 implemented in Python and enabled quality control, normalization of batch effects
3588 and biological interrogation of multiplexed SCP MS data. Our group developed
3589 UE-SCP (an Ultra-sensitive and Easy-to-use multiplexed Single-Cell Proteomic
3590 workflow) which was also inspired by SCoPE2 recently (Gu et al., 2022b). UE-SCP
3591 employed the cellenONE for sorting single cell softly and reduced the cell number in

carrier channel to 100 for better quantification. With the help of high sensitivity LC-timsTOF system, the median number of proteins identified from one HeLa cell can exceed 2,300, which achieved the largest identified protein number from one cell without any customized equipment up to now.

To avoid sample loss caused by nonspecific adsorption of tubes and LC columns, Li et al. brought a new idea called Mad-CASP (mass-adaptive coating-assisted single-cell proteomics) (Li et al., 2022g). They designed a hydrophobic peptide, which was mainly composed of hydrophobic amino acids (AAs) F and V with K inserted every 4 AAs (Figure 9B). Using tubes coated with these peptides to prepare single-cell samples, the number of identified proteins can increase by 63%. During trypsin digestion, these hydrophobic peptides could be digested into 5-AA peptide fragments. These low-mass fragments would be excluded in MS data acquisition and simultaneously play the role of carriers to reduce the loss of single-cell peptides due to the adsorption of the LC column. With this novel preparation and data acquisition strategy, they identified an average of 1240 proteins from a single HeLa cell.

Masuda et al. developed another novel SCP tool, called WinO (a water droplet-in-oil digestion) (Masuda et al., 2022). It was based on carboxyl-coated beads and phase transfer surfactants (Figure 9C). Single-cells were directly sorted into 96-well plates containing 50 μ l ethyl acetate and formed as a suspending droplet. Then the entire preparation was accomplished in this water droplet, minimizing the contact area between sample and containers to reduce the loss of proteins and peptides by adsorption. It was the first attempt to use magnetic beads in SCP to enhance the recovery of hydrophobic proteins and peptides. 96.2% of identified peptides showed higher intensity in samples prepared with the beads than in those without beads. This workflow has been successfully performed on 96-well plates and identified 845 proteins from one cell based on TMT10plex labelling.

Recently, Brunner et al. reported a true single-cell-derived proteomics (T-SCP) that aimed to combine the most advanced technologies that could achieve ultra-high sensitivity and be commercially available at the same time (Brunner et al., 2022). Cell sorting was achieved by FACS and the entire preparation happened in the commercial 384-well plate in microliter-level volume that was easy to operate. Then peptides were concentrated in an EvoTip device for desalt and on-line sample loading. With the help of timsTOF SCP MS and diaPASEF mode, 2,083 proteins were identified from one HeLa cell using a HeLa DIA spectral library with about 4,000 protein groups.

3627 Applications

Although the SCP technologies are not as developed as single-cell transcriptomics (SCT), there are still several studies that have demonstrated its indispensability in biological and clinical research. Here we mainly introduce its applications in cell differentiation, disease heterogeneity, and cell cycle.

3633 Cancer heterogeneity

Cell heterogeneity is an increasing concern in diseases, especially cancer research. Tsai et al. developed a novel SCP tool termed surfactant-assisted one-pot sample

3636 preparation (SOP)-MS and applied it to single luciferase 2-tdTomato (L2T) tumor
3637 cells derived from a patient CTC-derived xenograft (PCDX) mouse model, revealing
3638 different protein signatures between primary tumors and early lung metastases (Tsai
3639 et al., 2021). The differentially expressed proteins are involved in tumor immunity,
3640 epithelial cell differentiation and epithelial–mesenchymal transition (EMT), indicating
3641 the possibility of selective pressure in immune evasion and cell state plasticity. These
3642 results provide a clear path for future research into the mechanisms of cancer
3643 metastasis and have the potential to guide targeted cancer therapies
3644

3645 *Biomarker discovery*

3646 Mass spectrometry (MS)-based proteomics is an ideal tool for discovering differences
3647 in protein abundance levels in patients and healthy individuals, and therefore, in
3648 principle, a powerful technology for biomarker discovery. Ozge Karayel et al.
3649 (Karayel et al., 2022) performed large-scale cerebrospinal fluid proteomics analyses
3650 on Parkinson’s disease patients and quantified more than 1,700 proteins. They
3651 discovered lysosomal and immune-related biomarker signatures specific to
3652 Parkinson’s disease patients with LRRK2 G2019S carriers. Du et al. (Du et al., 2023)
3653 performed proteome profiling of 144 urinary and 44 urinary exosomes from type 2
3654 diabetes mellitus patients with albuminuria in varying degrees. By analyzing, they
3655 found several potential biomarkers, such as SERPINA1 and transferrin, that could be
3656 used for diabetic kidney disease diagnosis or disease monitoring.
3657

3658 *Cell differentiation*

3659 Cell differentiation processes are subject to various disturbances that lead to different
3660 cell fates. Single-cell proteome has been applied to reveal the heterogeneity and
3661 dynamics during cell differentiation. Using SCoPE-MS, Budnik et al. quantified the
3662 single-cell proteome of ES cell in days 3, 5, and 8 after differentiation induction
3663 (Budnik et al., 2018). They revealed the corresponding correlation vectors between
3664 days 5 and 8 were more similar than between days 3 and 5, indicating the more
3665 advanced differentiation changes on those days. Compared their SCP data with SCT
3666 data, they further found there is a coordinated mRNA and protein covariation at the
3667 single-cell level, proving the quantitative accuracy and necessity of SCP research.
3668 Using a multiplexed SCP workflow derived from SCoPE-MS, Schoof et al. explored
3669 the protein profiles of cells in different differentiative stages from a primary Acute
3670 Myeloid Leukemia (AML) culture model (Schoof et al., 2021). They successfully
3671 distinguished differentiation stages in this complex cellular hierarchy and found there
3672 might be two parallel differentiation trajectories for leukemic stem cells (LSC).
3673

3674 *Cell cycle*

3675 Analysis of the same cell type in different cell cycle phase is another challenge in
3676 single-cell omics which requires higher sensitivity and accuracy of detection. Brunner
3677 et al. applied T-SCP to demonstrate the protein profiles of HeLa cells which were
3678 arrested cell cycle by drug (Brunner et al., 2022). They investigated the differentially
3679 expressed proteins between different cell cycle stages, found a large number of known

3680 cell cycle regulators were significantly regulated, and also identified some new cell
 3681 cycle-associated proteins. Using the upgraded preparation method nPOP and LC-MS
 3682 analysis method pSCoPE, the research team of SCoPE2 recently quantified cell
 3683 division cycle (CDC)-related protein covariation within a cell type (Leduc et al.,
 3684 2022). They identified differentially expressed proteins among G1, S, and G2/M
 3685 phase in monocyte and melanoma cells, and constructed CDC markers.

3686

3687 *Discussion and prospects*

3688 At the genomic and transcriptomic levels, single-cell sequencing has become a
 3689 powerful tool for studying cell heterogeneity and identifying different phenotypic cell
 3690 types. In contrast, MS-based single-cell proteomics is still in its infancy. A recent
 3691 study by Brunner et al. sheds light on the necessity for single-cell proteomics
 3692 (Brunner et al., 2022). When comparing the SCP measurements with similar
 3693 single-cell RNA sequencing data, the protein expression completeness reached 49%
 3694 on average, whereas gene expression completeness was only 27% in SMART-seq2
 3695 and even as low as 8% for droplet-based data. Many of the transcripts are expressed at
 3696 less than one copy per cell on average and result in a mass of shot noise in SCT data.
 3697 Thus, the amount of transcriptional information that can be captured from one cell is
 3698 very limited, while single-cell proteomics can provide relatively complete information
 3699 on the protein level of one cell. This is particularly important for precious cell types
 3700 such as CTCs and embryos. On the other hand, several published works have
 3701 demonstrated a low correlation between SCT and SCP. Although bulk transcriptomic
 3702 and proteomic data showed a moderate correlation, the correlation at the single-cell
 3703 level decreased to 0.2~0.4 (Woo et al., 2021; Brunner et al., 2022). These results
 3704 illustrated that the proteome and transcriptome levels of the same gene can vary
 3705 greatly, but this difference was obscured in the bulk measurements. These views
 3706 further suggested the necessity for measuring proteome directly at the single-cell
 3707 level.

3708 The development direction of single-cell proteomics is always focused on more
 3709 identified proteins and higher throughput. Profit from both advances in instruments
 3710 and the development of new SCP workflows, the protein number identified from one
 3711 mammalian somatic cell has jumped to the level of 3,000. More than 5,000 proteins
 3712 can be cumulatively quantified in ~40 single-cells, which has been able to achieve a
 3713 similar data level as bulk-size proteomics (Wang et al., 2022d). By using the TMT
 3714 labeling strategy with shorter liquid chromatography gradients, more than 2,000
 3715 single cells have been completely analyzed in less than ten days, which has also been
 3716 close to matching the SCT throughput (Leduc et al., 2022). Although the deepest
 3717 protein coverage and the highest throughput have not been achieved in one SCP tool,
 3718 there are some methods that perform well in both. proteoCHIP achieved analysis of
 3719 more than 300 cells in about 2 days and identified 3,674 proteins from 276 single cells.
 3720 Our recent work, UE-SCP, identified 4,320 proteins from 128 single-cells and the
 3721 entire analysis can be completed within 3 days.

3722 At present, most SCP tools are limited to the laboratory where they were
 3723 developed and require highly specialized equipment or operator level. Ease of use is

an ineluctable issue for SCP to become a viable tool for scientific and clinical researches. There are several SCP tools that can perform excellently without any customized equipment, such as ScoPE2, UE-SCP, and T-SCP. At the same time, some highly integrated microfluidic chips showed promise for commercialization, such as proteoCHIP and N2 chips. Given that SCP is still in its early stages, most of the losses caused by the sample transfer process are still difficult to solve in unintegrated manners. We summarized the latest SCP tools from accessibility, throughput, and analysis depth shown in Figure S1. How to better balance these parameters is the next question to consider.

With the help of high-speed cell sorting machines, simplified sample preparation processes, and automated liquid operators, the steps prior to MS analysis have reached the throughput of more than 2,000 single cells per day (Leduc et al., 2022). However, under the premise of considering the analytical performance, the chromatographic separation and mass spectrometry detection time of one SCP sample is still about 1 h. LC-MS analysis time has become the bottleneck of SCP throughput improvement. Multiplex labeling is a feasible method to improve the throughput of LC-MS analysis which has been applied in many SCP tools yet. As TMT labels have been expanded to 18 plex, the development of higher plex reagents may require expensive investments but is still worth looking forward to. Nonisobaric isotopologous mass tags such as mTRAQ have been used in low-input proteomics and combined with DIA-MS to improve throughput and protein coverage at the same time (Derks et al., 2022). As nonisobaric isotopologous mass tags may be easier to expand to multiplex than isobaric mass tags, its application in high-throughput single-cell proteomics is promising. Dephoure and Gygi described a hyperplexing method that enabled the analysis of samples from multiple conditions simultaneously by combining two different labeling methods, which may have implications for throughput enhancement in single-cell proteomics (Dephoure and Gygi, 2012).

Here we focused on proteomics based on living single-cells in suspensions, but it is important to note that SCP based on trace cells from formalin-fixed paraffin-embedding (FFPE) tissues has come hand in hand. Although it is not yet possible to identify thousands of proteins from a single cell in FFPE tissues because of the impact of formaldehyde-mediated cross-links, there are several approaches that are approaching this goal. Using the optimized nanoPOTS, Nwosu et al. have identified an average of 1,312 from mouse liver tissues as small as $0.0025 \text{ mm}^2 \times 10 \mu\text{m}$ which corresponded to about 10 cells (Nwosu et al., 2022). We developed a spatially resolved proteomic tool called LCM-MTA which can quantify 536 proteins from $0.005 \text{ mm}^2 \times 8 \mu\text{m}$ human placenta FFPE tissue (about 15 cells) and 1,477 proteins from $0.1 \text{ mm}^2 \times 8 \mu\text{m}$ tissue (Gu et al., 2022a). Applied the LCM-MTA on clinical colorectal cancer (CRC) tissues, the functional differences of different cell types were accurately distinguished. Mund et al. introduced the Deep Visual Proteomics (DVP), which combined artificial-intelligence-driven image analysis of cellular phenotypes with automated single-cell or single-nucleus laser microdissection and ultra-high-sensitivity mass spectrometry. By collecting about 100 cells for one sample, they have successfully characterized the expression of the proteome from

3768 melanocytes, melanoma in situ to invasive melanoma (Mund et al., 2022b). Spatially
 3769 proteomics in single-cell resolution can provide a new dimension to single-cell
 3770 proteomics. Many hospitals and research institutions have massive amounts of FFPE
 3771 tissue stored in their repositories. If reliable SCP tools can be applied to them, it will
 3772 bring a great boost to biomedical research.

3773

3774 *Summary*

3775 Overall, single-cell proteomics is in the early stage of explosive development. Just in
 3776 2019, analysis of the proteome from single cells was described as a “dream”, but
 3777 today there have been several promising tools developed (Marx, 2019). We believe
 3778 that with the optimization of accessibility and the further improvement of throughput,
 3779 the truly large-scale applications of single-cell proteomics in scientific and clinical
 3780 research, such as organ maps, drug screening, and precise disease classification, are
 3781 within reach.

3782

3783 **Chapter 5 Single-cell metabolomics technology**

3784

3785 Most human cells are approximately 5 to 25 μm in diameter with as low as sub-pl
 3786 intracellular volumes and highly dynamic metabolite concentrations ranging from a
 3787 few copies to more than 100,000 (Zenobi, 2013). Compared with other omics, the
 3788 genome is approximately static, the proteome and transcriptome change in minutes or
 3789 hours, whereas the metabolome changes on a time scale of milliseconds to seconds
 3790 (Weibel et al., 1974). For the present objects, the metabolome includes small
 3791 molecules (usually lesser than 1.5 kDa in size, but excluding nucleic acids, minerals,
 3792 and salts), lipids, peptides, drugs, and their xenobiotics (Wishart et al., 2007;
 3793 Minakshi et al., 2019). All of them are characterized by structural diversity, which
 3794 makes discrimination difficult. And typically, a single cell can be detected tens to
 3795 hundreds of analytes but only ~10% can be assigned using high-resolution mass
 3796 spectrometry (MS) and database search methods (Yin et al., 2018). Thus, extracting
 3797 small volume content, snapshotting the quick turnover of metabolites, discriminating
 3798 the molecular species diversity, improving detection sensitivity and selectivity, and
 3799 boosting detection limits, all are inevitable challenges in single-cell metabolomics
 3800 (SCM) research.

3801

In the process of SCM research, a large number of research technologies,
 analytical platforms, and applications have emerged. Here, we review the
 development of SCM in the last ten years, including the classes of research techniques,
 mainly analytical workflow, applications, and possible breakthroughs.

3805

3806 *Research techniques in single-cell metabolomics*

3807

To date, there are various research techniques for analyte measurement of a single cell,
 which are mainly divided into microscope-based, spectroscopy-based, and mass
 spectrometry-based platforms (Galler et al., 2014). Microscope-based technologies
 could observe cellular structures at the nanoscale, such as stimulated emission
 depletion (STED) microscopy, stochastic optical reconstruction microscopy

(STORM), photoactivated localization microscopy (PALM), and so on. The advantages of microscope-based analysis for single cells are as obvious as the disadvantages, with the highest spatial-resolution insight into cellular structures but the least biochemical information. Furthermore, most microscope-based methods have a low throughput limitation, and their long detection time is not suitable for dynamic analysis (Zheng and Li, 2012). Spectroscopy-based methods are widely applied, among which nuclear magnetic resonance (NMR) is the most used because of its characteristics of nondestructive detection and high reproducibility. However, multicellular analysis has dominated so far due to its relatively low sensitivity (Galler et al., 2014). Mass spectrometry-based methods are the indispensable tool for the simultaneous detection of a large number of analytes in a short period of time. They provide accurate mass-charge ratios, retention times, and quantitative results for both known and unknown molecules. In addition, molecules below sub-attomolar concentrations could be detected (Villas-Bôas et al., 2005). By contrast, MS wins out among these technologies for its high detection sensitivity and selectivity, broad detection range, fast analysis speed, and strong power of molecular structure identification. MS is considered the most powerful tool for characterizing the chemical profile of a single cell.

3830

3831 *State-of-the-art technologies and methods in single-cell metabolomics field based on*
3832 *mass spectrometry*

3833 The analytical workflow in single-cell metabolomics based on MS mainly refers to
3834 single-cell sampling, content measurement, and data analysis.-Single-cell sampling is
3835 the core of SCM. Single cells can be sampled directly or cultured on other platforms
3836 until metabolite analysis. For the purpose of the sampled content truly reflecting the
3837 metabolic profile (for example, neither loss of analyte volume nor misleading results
3838 caused by rapid metabolic turnover), sometimes additional treatments are needed to
3839 quench cell metabolism, including the addition of cold organic solvents or rapid
3840 freezing (Ibáñez et al., 2013). Care must be taken to avoid interfering with the culture
3841 media which may lead to the production of abnormal metabolites (Minakshi et al.,
3842 2019). During the content measurement by MS, molecules are ionized and converted
3843 to the gas phase, followed by passed into MS. Then moving ions are separated
3844 according to their mass to charge ratios in the magnetic field or electric field and
3845 detected by a detector. There are two main types of ion sources have been applied in
3846 single-cell metabolomics, laser desorption ionization and electrospray ionization
3847 (Figure 10). However, data acquisition could be challenging and the mode needs to be
3848 selected according to the research purpose. A large but indistinguishable number of
3849 metabolite features would be obtained if untargeted metabolomics is performed,
3850 whereas limited but definitive results would be acquired if targeted metabolomics is
3851 performed. Therefore, it is necessary to consider the tradeoff between throughput and
3852 coverage before SCM analysis (Tajik et al., 2022). In order to obtain the structural and
3853 functional information of metabolites, data analysis is carried out. The process of
3854 information mining partly determines the results of the research, which is important

3855 for the research. Therefore, we focus on single-cell sampling techniques and data
 3856 analysis.

3857

3858 *Sampling techniques in single-cell metabolomics based on Mass spectrometry*

3859 The single-cell sampling techniques for MS analysis can be broadly divided into three
 3860 categories: (1) desorption ionization, (2) content extraction, and (3) sorting and
 3861 ionization.

3862

3863 *Desorption ionization*

3864 The intact single cells can be directly subjected to MS analysis where sampling and
 3865 ionization processes simultaneously occur using the desorption ionization method
 3866 (Liu and Yang, 2021). Desorption ionization can be divided into vacuum desorption
 3867 and ambient desorption based on whether analytes are ionized in vacuum system.

3868 Secondary ion mass spectrometry (SIMS) (Figure 10A) uses a high-energy
 3869 accelerated primary ion beam (e.g., Cs^+ , O_2^+ , Ar^+ , and Ga^+) to bombard the target
 3870 surface, which results in the ejection of plume of molecules or ions from the surface.
 3871 SIMS is an effective technique for subcellular distribution imaging of various
 3872 molecules with high spatial resolution (50 nm) (Yin et al., 2019a). While traditional
 3873 primary ion beams induce extensive molecular fragmentation, modern SIMS often use
 3874 cluster ions as its primary ion beam (e.g., Bi_3^+ , SF_5^+ , and C_{60}^+) to minimize
 3875 fragmentation (Rubakhin et al., 2013). Nanoscale secondary ion MS (nanoSIMS)
 3876 enables the primary ion beam to scan the sample at a perpendicular angle which
 3877 shortens the working distance and improves secondary ion transmission. It has been
 3878 applied for quantitation of subcellular chemical distribution with a lateral resolution
 3879 of ~50 nm (Jiang et al., 2014).

3880 Matrix-assisted laser desorption ionization (MALDI) (Figure 10B) is regarded as
 3881 a soft ionization method that does not cause excessive fragmentation. Since the
 3882 increasing ionization efficiency, it has become one of the most widely used laser
 3883 desorption methods, which relies on the absorption of laser and the transfer of charges
 3884 by auxiliary matrix molecules to enhance the analyte ionization. It can achieve
 3885 resolution at the micron to submicron level (Emara et al., 2017) and provide
 3886 high-fidelity results of native analyte distribution.

3887 However, signals from the matrix molecules may strongly overlap with potential
 3888 analytes (<500 Da) (Ferguson et al., 2014). Multiple matrix-free laser desorption
 3889 ionization (LDI) methods, such as desorption/ionization on porous silicon (DIOS) MS,
 3890 Nanostructure-initiator MS (NIMS), and nanopost array (NAPA) MS (Figure 10C)
 3891 were used. These ionizations rely on the interaction between laser radiation and
 3892 nanostructures to contribute to the desorption and ionization of the sample, which
 3893 solve the signal overlap and have comparable spatial resolution to MALDI (Yin et al.,
 3894 2019a).

3895 Traditional SIMS, MALDI, and matrix-free LDI methods are all vacuum-based.
 3896 Qualitative and quantitative information on small molecular substances can be
 3897 provided due to excellent spatial resolution. The optimized technology has high
 3898 sensitivity and detection limits can reach the fg level (Yin et al., 2019a). A few

thousand cells can be analyzed in a single experiment after generating single-cell arrays (Zhang and Vertes, 2018). However, some sample preparation steps, such as frozen dehydration, are introduced to maintain cellular shapes under high vacuum conditions (Zhang and Vertes, 2018). These steps may affect the chemical compositions of cells. The vacuum condition can also potentially interfere with the distribution of metabolites, particularly for volatile and semi-volatile species. Owing to technical innovations, atmospheric-pressure MALDI-MS (AP-MALDI-MS) has been developed with a lateral resolution as low as 1.4 μm (Kompauer et al., 2017).

Ambient ionization refers to the generation of ions under ambient conditions (e.g., native temperature and pressure) requiring little to no sample preparation. Laser ablation electrospray ionization (LA-ESI) (Figure 10D) is a matrix-free technique that utilizes a pulsed mid-infrared region laser beam at the wavelength of 2.94 μm to activate a water-rich target sample. At this wavelength, water strongly absorbs the laser radiation and creates a plume of molecules that are released into the gas phase (Nemes and Vertes, 2007). The desorption plume mixes with an ESI plume can enhance the ionization of analytes. LA-ESI eliminates sample preparation and has a spatial resolution as low as 30 μm (Shrestha and Vertes, 2009). A similar approach is desorption electrospray ionization (DESI) (Figure 10E), in which analyte ions are produced by desorption and ionization using electrospray directed toward the sample surface. However, the limited spatial resolution ($>50 \mu\text{m}$) usually prevents it from SCM analysis (Taylor et al., 2021). Nanospray desorption electrospray ionization (nano-DESI) utilizes a primary capillary for solvent delivery on cell samples and a secondary capillary for picking up the extracted molecules for MS analysis. Its resolution is determined by the size of the liquid bridge formed between two capillaries, which is controlled by the capillary's size, their position, and the flow rate (Yin et al., 2019b). The addition of shear force probes standardizes capillary-to-sample distance (Nguyen et al., 2017). Then, a pneumatically assisted nano-DESI device was implemented to propel the solvent through the nanospray capillary, which improved sensitivity for metabolite species by 1~3 orders of magnitude and reduced ionization suppression (Duncan et al., 2017). These reduce the dependence on probe-to-surface distance. The resolution that can be achieved with current nano-DESI technology is 8.5 μm (Rao et al., 2015).

The sensitivity, spatial resolution, coverage, and throughput vary with the different desorption ionization methods (Figure S2 and Table S8) (Taylor et al., 2021). At present, desorption ionization MS methods are mainly used in mass spectrometry imaging (MSI) which employs an analytical probe (e.g., ion beam, laser, and solvent junction) capable of analytes desorption and ionization *in situ*. MSI can provide additional functional information by mapping the location of small molecules *in situ*, which is promising (Taylor et al., 2021). SpaceM integrated MALDI imaging with light microscopy and digital image processing. It took the first microscope image to capture the relative positions of cells, then collected MALDI imaging of metabolites, followed by a second microscope image to show a visual cue which cell the metabolite came from. SpaceM could detect >100 metabolites from $>1,000$ individual cells per hour (Rappez et al., 2021). It has the most identifications among the known

3943 MSI techniques. If the target sample is sectioned consecutively and each section is
 3944 used for MSI, the 3D spatial metabolite map will be obtained after the compilation of
 3945 the 2D MS images. Dueñas et al. utilized MALDI-MSI to visualize the
 3946 three-dimensional spatial distribution of phospholipid classes in individual zebrafish
 3947 embryos (Dueñas et al., 2017).

3948

3949 *Content extraction*

3950 Electrospray ionization methods (Figure 10F) are also extensively used for biological
 3951 molecule analysis of single cells, especially for live cells, because of the significantly
 3952 reduced mechanical and chemical perturbations and greatly simplified sample
 3953 preparation. ESI-MS generally favors the detection of analytes at relatively high
 3954 concentrations due to its relatively low ionization efficiency, ion transmission, and
 3955 relatively high ion suppression. Naturally, the modified techniques have been
 3956 developed and applied in SCM, including nano-electrospray ionization (nano-ESI)
 3957 (Figure 10G) (Karas et al., 2000), probe electrospray ionization (PESI) (Gong et al.,
 3958 2014), induced nano-ESI (InESI) (Huang et al., 2011) and pico-electrospray
 3959 ionization (pico-ESI) (Wei et al., 2020) (Figure 10H).

3960 Metabolome acquired directly from a living cell *in situ* can result in more
 3961 realistic and representative chemical profile of cell metabolism and phenotype. The
 3962 direct sampling analysis method is content extraction which can be divided into
 3963 micromanipulation, microextraction, and microjunction probes (Figure 11).

3964 Micromanipulation means manipulating a micropipette to gently pick individual
 3965 cells and suck out metabolites. Micromanipulation coupled MS mainly uses nano-ESI
 3966 capillary whose emitter internal diameter is closer to the MS inlet. An application,
 3967 known as live single-cell MS (live MS) (Figure 11A), was achieved by sucking out
 3968 the content with a metal-coated microcapillary under video-microscopy, adding an
 3969 ionization solvent (acetonitrile containing 0.5% formic acid) from the microcapillary
 3970 rear, and directly feeding the mixture into MS (Mizuno et al., 2008). The extracted
 3971 content characterized hundreds of molecules at sub-attomolar-level sensitivity within
 3972 minutes (Fujii et al., 2015). However, the analytes were diluted tens of thousands of
 3973 times due to the ionization reagent. Subsequently, PESI was used to enrich and extract
 3974 metabolites by inserting a tungsten probe with a tip diameter of 1 μm into the single
 3975 cell for ~30 s (Gong et al., 2014). Both have the disadvantage of controlling the
 3976 imprecise amount of extracted material from cells. Consequently, quantitative
 3977 extraction techniques of pressure assisted (Zhang and Vertes, 2015) or electroosmotic
 3978 (Yin et al., 2018) microsampling were developed.

3979 Most small molecules from a single cell can't be directly detected by MS due to
 3980 the presence of intracellular interfering ions and high concentrations of non-volatile
 3981 salts. Thus, liquid-liquid extraction serves different analyte classes (Figure 11B).
 3982 Multiple microextraction devices coupled with MS are proposed to achieve a high
 3983 coverage metabolic analysis by adding low volumes of an extraction solvent. In short,
 3984 the capillary tip absorbs organic solvent and aqueous solution respectively in positive
 3985 and negative ion mode at nano-liter or pico-liter scale, followed by connected to a
 3986 syringe and a clamp which enable extraction reagent to cover a single cell for a few

seconds (e.g., 10 s) under an inverted microscope and to complete the extraction (Wang et al., 2019c). In general, the microextraction partially alleviates the problem of MS incompatibility with intracellular interfering ions and salts, but the high viscosity of the cell contents also needs assistant ionization solvents to obtain ion signals which limits metabolite coverage.

The characteristics of the above content extraction, sampling and then ionization, prohibit the real-time detection. Liquid microjunction probes achieve in real-time, *in situ* metabolite extraction with integrated solvent microextraction and nanoscale micromanipulators. Yang's group introduced multiple devices, including Single-probe (Figure 11C) and T-probe. Single-probe is fabricated by embedding a fused silica capillary and a nano-ESI emitter into a dual-bore quartz needle. Droplets at the tip of the needle formed by continuously injecting liquid from one side of the needle can extract intracellular material and they are discharged through the other side and sent to MS (Pan et al., 2014). T-probe works similarly, with three capillaries embedded into T-shaped grooves, where the solvent-providing capillary is in line with the nano-ESI emitter and the sampling capillary is vertically placed (Liu et al., 2018b). Single-probe is used in the single cells residing in microwells and T-probe is modified to analyze live non-adherent cells. In terms of operation difficulty, the content extraction of adhesive cells is easier to achieve than suspended cells (Emara et al., 2017).

In brief, the method of content extraction can not only realize living cell studies but also obtain relatively high metabolite coverage. However, the main limitations of the content extraction method are relatively tedious manipulation, low sample throughput, and time-consuming (3~5 min per cell) (Table S8) (Fujii et al., 2015).

4011

4012 *Sorting and ionization*

4013 Most of the time, single-cell analysis begins with sample preparation in a bid to
 4014 isolate the target cell without affecting its state. There are a large number of
 4015 single-cell isolation techniques in a high-throughput manner that have been developed
 4016 so far. Since single-cell sampling occurs after sorting, we call this method sorting and
 4017 ionization for short. It can be divided into label sorting-based, microfluidic
 4018 device-based and laser capture microdissection (LCM)-based methods (Figure 12).

4019 Conventional cell sorting methods, including flow cytometry,
 4020 fluorescence-activated cell sorting (FACS), and mass cytometry, are label-based. Flow
 4021 cytometry flows a cell at a time by controlling the cell suspension. Labeled with
 4022 fluorescent markers, cells can shed light into various properties when a laser beam
 4023 scatters through them. Similarly, FACS also uses light scattering and fluorescence
 4024 properties to sort cells into subpopulations. As a combination of flow cytometry and
 4025 MS, mass cytometry, of which antibodies are labeled with heavy metal ion tags
 4026 instead and detected by inductively coupled plasma (ICP), has been used for sorting
 4027 and targeted high-throughput molecular analysis (Bandura et al., 2009). Mass
 4028 cytometry is currently limited to a couple of dozen available proteins and is not shown
 4029 in metabolite analysis because the small molecules are difficult to label. Nonetheless,
 4030 a label-free mass cytometry realizes online sorting and real-time ESI-MS analysis for

4031 a single cell. CyESI-MS uses three coaxial capillaries to deliver cell suspension,
 4032 sheath fluid, and sheath gas, respectively (Figure 12A). Cells are isolated and
 4033 extracted by the sheath fluid, then the sheath gas assists solvent evaporation and
 4034 ensures the ions enter MS, which could simultaneously detect hundreds of cellular
 4035 metabolites in a high-throughput way, approximately 38 cells min⁻¹ (Yao et al., 2019).

4036 Microfluidic devices bring a significant enhancement in the throughput and
 4037 simplification of the workflow. They physically confine individual cells to
 4038 microfluidic structures, among which micro/nano-wells, droplets,
 4039 microvalve-controlled channels, and hybrid microfluidic platforms are most
 4040 extensively used in single-cell analysis (Liu et al., 2019).

4041 Micro/nano-well-based microfluidic devices, also known as chip-based methods,
 4042 consist of dense arrays of wells that typically are lithographically fabricated onto
 4043 polydimethylsiloxane (PDMS), glass, or silicon and serve as containers for individual
 4044 cells (Torres et al., 2014). For example, the early invention of an integrated
 4045 microfluidic array plate (iMAP) was characterized by the interface of gravity driven
 4046 flow, open input fluid exchange and cell capture mechanism with approximately 100%
 4047 capture rate (Dimov et al., 2011). Its design allowed for single-cell capture, reagent
 4048 addition, and parallel processing operations. Castro and coworkers deposited a small
 4049 volume of buffer containing dense-core vesicles and electron lucent vesicles of
 4050 *Aplysia californica* cells onto an indium tin oxide (ITO)-coated glass slide (Castro et
 4051 al., 2021). Ibanez and colleagues developed microarrays for MS (MAMS) (Figure
 4052 12B) that allowed thousands of individual cells to be analyzed in a single MS
 4053 experiment, which featured arrays of hydrophilic wells patterned on an omniphobic
 4054 surface to enable automated isolation of single cells (Ibáñez et al., 2013). Yang and
 4055 coworkers revised the fabrication process of the microdot array by using the contact
 4056 printing technique (Yang et al., 2016).

4057 Droplet-based microfluidic devices usually use two immiscible fluids to create
 4058 water-in-oil micro/nanodroplets containing the individual cell as single-cell reaction
 4059 vessel. A single cell in the droplets could be achieved by limited dilution, but the
 4060 probability of single-cell events is limited. Combined an inkjet nozzle cell
 4061 manipulator with PESI-MS, a drop-on-demand inkjet printing device was fabricated
 4062 and used for lipid profiling, which was capable of producing single-cell events with a
 4063 probability of about 50% in a fully automatic manner (Chen et al., 2016a). In the other
 4064 example, Lin's group designed a Dean flow assisted cell ordering system (Figure 12C)
 4065 to detect multiple cellular lipids, of which a spiral capillary generated a secondary
 4066 force to separate particles in a single equilibrium position and reduce the
 4067 agglomeration (Huang et al., 2018b).

4068 Microvalve-based devices utilize parallel microchannel circuits coupled to
 4069 pressure-controlled valves or similar control devices (Figure 12D) (Unger et al., 2000).
 4070 By means of precise control of the microvalve assembly, a series of operations
 4071 automate and parallelize complex biological analysis.

4072 The microfluidic platforms aforementioned have their own advantages and
 4073 drawbacks. A promising approach is to combine the core components from different
 4074 microfluidic platforms and overcome each other's limitations to form a new hybrid

platform, whose common name is lab-on-a-chip system. Leung and coworkers developed a microfluidic device. The programmable microdroplets-based device combined integrated microvalve technology with the sample compartmentalization and dispersion-free transport to perform single-cell manipulation and analysis (Leung et al., 2012). Furthermore, a multi-dimensional organic mass cytometry was established by connecting a simple microfluidic chip to the nanoelectrospray emitter, enabling the identification of about 100 metabolites with a throughput of around 40 cells min⁻¹ (Xu et al., 2021a) (Table S8). The multi-step integration is not only beneficial to obtain high coverage results but also to save time and labor.

Sorting and selection of single cells or subcellular components can be done by the LCM system which typically consists of a microscopy component for sample visualization, a laser component for selectively dissecting samples, and a collection component for material dissected. It is not considered a high-throughput method, but a high-resolution method suitable for single-cell isolation from tissue section samples. LCM can completely isolate the cell from its natural environment. There has developed a high spatial resolution hybrid laser capture microdissection/liquid vortex capture/mass spectrometry system (LMD-LVC-MS) (Figure 12E) with a liquid vortex capture probe placing directly below the sample substrate that captured the laser-ablated material, dissolved the material into liquid, and transported it to MS for analysis (Cahill and Kertesz, 2018).

It is worth mentioning that although the above sampling techniques can be applied to live cells, developing technologies under near-physiological conditions remains challenges. But there have been some technological breakthroughs. Shao et al. came up with an intact living-cell electrolaunching ionization mass spectrometry (ILCEI-MS) method, which used a capillary with an inner diameter slightly smaller than the average cell diameter to achieve cell separation and transport with the help of a small device (Shao et al., 2022b). It reduced the volume of ionized droplets formed under the combined action of an applied electric field and the surface tension on the port before reaching the MS inlet. A droplet was roughly equal to a cell. Through this method, 51 cells could be analyzed per minute, and 368 metabolites (from 482 cells) could be assigned in a single experiment. Recently, an asymmetric serpentine channel microfluidic chip coupled to pulsed electric field – induced electrospray ionization – high resolution mass spectrometry (chip-PEF-ESI-HRMS) conditions have been developed, which was sheathless and external-force-unused. The single cells were suspended in an aqueous solution (i.e., isotonic salt concentration). Once a single cell reached the tip of nanospary emitter, the high voltage electric field made it disruption and the contents ionization and identification in real time. It allowed for the annotation of approximately 120 metabolites in a single cell and the throughput of up to 80 cells min⁻¹ (Feng et al., 2022).

4114

4115 *Data analysis in single-cell metabolomics based on Mass spectrometry*

4116 The workflow of current single-cell metabolomics data analysis includes data
 4117 preprocessing, metabolite annotation, statistical analysis, network and pathway
 4118 analysis, and data visualization. A large number of bioinformatics tools, analytical

software, and databases are now available at each step (Misra, 2020; Liu and Yang, 2021) (Table S9) Data preprocessing includes two parts. First, convert the raw data only available to the commercial software of specific vendor into a compatible format. Second, extract metabolome related information which involves determining “true” signals, normalizing relative abundances among different cells, and screening out the metabolites present in most cells. Except for the methods mentioned in Table S9, more and more custom algorithms have been written to extract information (Liu and Yang, 2021; Shao et al., 2022b). Then some software is needed to recognize metabolites, which is metabolite annotation. Next is statistical analysis. In order to reduce batch effects and technical variations, data processing is carried out, including normalization, transformation, and scaling. After evaluating whether the data has Gaussian distribution or not, parametric or nonparametric univariate analysis (e.g., t-tests and Analysis of Variance) and multivariate analysis (e.g., unsupervised principal components analysis and supervised orthogonal partial least-squares discriminant analysis) are performed to reveal the metabolomic biomarkers, group clustering results, and discrimination between groups according to the experimental design. By integrating conventional statistical methods with machine learning to build complex mathematical models, it can provide high predictive classification results. The open-source statistical analysis tools include, but are not limited to, R (<http://www.R-project.org>) and Python (<https://www.python.org>). Mapping of metabolites onto metabolic maps or known biochemical pathways is network and pathway analysis. Finally, the above results are visualized to complete the data analysis.

4142

4143 *Applications*

4144 Subtle differences from cell to cell may lead to great changes in important biological
 4145 processes. The bulk analysis of cell population shows the average features of multiple
 4146 cell types and ignores the rare cells. Therefore, it is necessary to study individual cells.
 4147 At present, SCM has been applied to various studies and has made more or less
 4148 progress.

4149 It is commonly used to identify single-cell metabolites, explore the metabolic
 4150 profiles and compare the up-regulation and down-regulation of metabolites in normal
 4151 and other states, especially for cancer cells. It is also applied to quantify compounds
 4152 (Pan et al., 2019), visualize cell heterogeneity (Huang et al., 2018b), differentiate cell
 4153 subsets (Zhang et al., 2018), and screen drugs (Anchang et al., 2018). For example,
 4154 abnormal lipid synthesis (e.g., C = C bond position or sn-position isomers formation)
 4155 could result in different diseases and reflect the prospect of lipidomics in precision
 4156 medicine (Li et al., 2021d). It served to investigate the role of cells in key biological
 4157 processes, including drug resistance (Prieto-Vila et al., 2019), immune response
 4158 (Labib and Kelley, 2020), tumor metastasis (Wu et al., 2020), and cell fate
 4159 determination (Stirparo et al., 2018). Also, an important application is to study the
 4160 special properties of rare cells, such as circulating tumor cells (CTCs), cancer stem
 4161 cells, and antigen-specific T cells. Take CTCs for example, they are released into the
 4162 bloodstream from primary tumor lesions and cause metastases in distant tissues and

4163 organs (Abouleila et al., 2019). Abouleila et al. revealed the metabolic profile
 4164 differences between CTCs and lymphocytes and found that the synthesis of GPLs was
 4165 a key factor in cancer proliferation. Moreover, SCM is also used in plants. For
 4166 example, Yuan et al. used metabolomics-assisted breeding in watermelons (Yuan et al.,
 4167 2023).

4168

4169 *Summary*

4170 Current single-cell metabolomics methods can analyze a maximum of about 3,000
 4171 cells in a single experiment, mostly 500~1,000 (Feng et al., 2022). Developing
 4172 automation technology is beneficial. For example, armed with an automated system
 4173 and a recognition algorithm, a dispenser robot coupled to a motorized x-y stage
 4174 enables to pick up the target cell or organelle quickly (Emara et al., 2017).

4175 The destructive nature of MS limits the repeatability and temporal analysis.
 4176 Therefore, a combination of different analytical tools is recommended to provide
 4177 comprehensive and multi-dimensional information. For example, a modified patch
 4178 clamp setup was combined with InESI-MS to simultaneously capture the
 4179 electrophysiological and metabolic state of a neuron (Zhu et al., 2017). At the same
 4180 time, it is necessary to integrate with single-cell multiomics to plot more detailed
 4181 characteristic profiles.

4182 How to get highly confident results is a much more important thing than crafts.
 4183 Since the standardization procedure of single-cell metabolomics, from sample
 4184 preparation to data analysis, has not been established, the following work should
 4185 follow it. We summarize the technological progress over the last decade in the Table
 4186 S8. Balancing sensitivity, coverage, and throughput, none of these methods are perfect.
 4187 So, it is necessary to develop new methodologies to improve all aspects. In summary,
 4188 SCM is still in the early stage, it is going to continue to flourish.

4189

4190 **Chapter 6 Single-cell multimodal sequencing technology**

4191

4192 Multiomics analysis with bulk sequencing has been widely applied to provide a
 4193 comprehensive understanding of biological processes like disease development
 4194 (Hoadley et al., 2018; Hutter and Zenklusen, 2018; Liu et al., 2018a; Malta et al.,
 4195 2018) and tissue development (Consortium et al., 2020a; Consortium et al., 2020b; He
 4196 et al., 2020b; Sethi et al., 2020) through the atlas integration of multi-omics datasets
 4197 like genomic, transcriptomic, epigenomic and proteomic data in multiple species. At
 4198 the single-cell scale, the applications of unpaired single-cell multi-omic sequencing
 4199 technologies, which use different cells from the same or similar source for different
 4200 single-cell experiments, have been applied to discover new cell subpopulations and
 4201 new biological mechanisms by connecting the single-cell transcriptome for cell type
 4202 identification to different modalities of other similar cells (Argelaguet et al., 2019;
 4203 Hao et al., 2021b). Recent advances in single-cell sequencing technologies have
 4204 further enabled the measurement of multiple omics like DNA, mRNA, epigenomic,
 4205 and protein in the same cell at single-cell resolution, providing the paired and
 4206 high-resolution discovery of single-cell status. Also, to integrate the multi-omic

4207 single-cell datasets, several bioinformatic algorithms and methods have been
 4208 developed to help pre-process, integrate, and interpret the emerging multi-omic
 4209 single-cell datasets.

4210 In this chapter, we first review the 1) recent development of single-cell joint
 4211 profiling technique capturing multiple views of cell molecules from the same cell,
 4212 mainly focused on the transcriptome-focused multimodal technologies (Table 3). Next,
 4213 we review the 2) recent advances in multi-modal integration analysis methods and
 4214 tools, including the different categories of applications and algorithms for both
 4215 unpaired multimodal datasets and paired multimodal datasets. Also, we summarize the
 4216 performance of popular single multi-omic data integration methods from recent
 4217 single-cell multi-modal integration benchmark studies.

4218

4219 *Single-cell multimodal sequencing technology*

4220 Multiple types of molecules can be isolated from the same captured cell by single-cell
 4221 multi-omics technologies. Several approaches of single-cell multi-omics sequencing
 4222 designed for capturing genomic DNA (gDNA), transcriptome, proteome, and
 4223 epigenome have been developed in recent years. Major steps of single-cell
 4224 multimodal sequencing technology workflow as depicted in Figure 13.

4225

4226 *Transcriptome + gDNA*

4227 Several multi-omic technologies can simultaneously measure mRNA and gDNA in a
 4228 single cell. Genome and transcriptome sequencing (G&T-seq) (Macaulay et al., 2015)
 4229 applied flow cytometry cell isolation along with beads-based mRNA and gDNA
 4230 separation. gDNA-mRNA sequencing (DR-seq) (Dey et al., 2015) isolated cells by
 4231 pipette and then amplified and split tagged gDNA and mRNA. Simultaneous
 4232 isolation of genomic DNA and total RNA (SIDR) (Han et al., 2018a) selected cells
 4233 with microplates and separated nucleus and cytoplasm by hypotonic cytolysis. And
 4234 TARGET-seq (Rodriguez-Meira et al., 2019) optimized the steps of
 4235 fluorescence-activated cell sorting (FACS) for cell isolation and reverse transcription
 4236 polymerase chain reaction (RT-PCR) for amplification, which provided higher cell
 4237 throughput than previous methods.

4238

4239 *Transcriptome + epigenome*

4240 Bisulfite (BS) treatment can convert methylated and unmethylated DNA CG sites
 4241 (Frommer et al., 1992) and analyze the DNA methylation at single nucleotide
 4242 resolution by PCR and next generation sequencing (Grunau et al., 2001; Harris et al.,
 4243 2010). Several single-cell bisulfite sequencing methods that measure the methylation
 4244 level at single-cell scale have been developed, including single-cell reduced
 4245 representative bisulfite sequencing (scRRBS) (Guo et al., 2013), single-cell whole
 4246 genome bisulfite sequencing (scWGBS) (Smallwood et al., 2014), single-nucleus
 4247 methylcytosine sequencing (snmC-seq) (Luo et al., 2017), and single-cell
 4248 combinatorial indexing for methylation (sci-MET) (Mulqueen et al., 2018) (see
 4249 “Epigenome sequencing” section for more detail). Recently, several single-cell
 4250 multi-omic techniques have been developed to capture mRNA and gDNA methylation

in the same cell. Firstly, single-cell methylome and transcriptome sequencing (scM&T-seq) (Angermueller et al., 2016) used similar protocol as G&T-seq (Macaulay et al., 2015), which used flow cytometry cell isolation along with beads-based mRNA and gDNA separation followed by bisulfite treatment. Secondly, simultaneous single-cell methylome and transcriptome sequencing (scMT-seq) (Hu et al., 2016b) used micro pipetting to isolate the nucleus of the single cells, then performed scRRBS and a modified Smart-seq2 procedure to generate DNA methylome and transcriptome data, respectively. As an extension of scMT-seq (Hu et al., 2016b), scTrio-Seq (Hou et al., 2016) can analyze genomic CNVs, the DNA methylome, and the transcriptome for individual cell simultaneously, as the genomic CNVs can be computationally inferred from scRRBS by bulk RRBS data.

Several next-generation-sequencing-based techniques, such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Johnson et al., 2007; Blecher-Gonen et al., 2013), Dnase I hypersensitive site sequencing (Dnase-seq) (Boyle et al., 2008), and assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013) have been developed to investigate the epigenome profiles such as chromatin structure and histone modifications in many species (Consortium et al., 2020a). Another similar method, nucleosome occupancy and methylome sequencing (NOME-seq) (Kelly et al., 2012) can label accessible genomic regions using an exogenous M. CviPI GpC methyltransferase and simultaneously measure nucleosome occupancy and cytosine methylation level. Based on these methods, many new protocols have been developed to measure the chromatin accessibility as well as DNA methylation or histone modification in chromatin accessible sites at single-cell resolution, including single-cell Dnase sequencing (scDNase-seq) (Jin et al., 2015), single-cell combinatorial indexing assay for transposase-accessible chromatin with sequencing (sci-ATAC-seq) (Cusanovich et al., 2015), single-cell assay for transposase-accessible chromatin using sequencing (sc ATAC-seq) (Buenrostro et al., 2015), single-cell micrococcal nuclease sequencing (scMNase-seq) (Lai et al., 2018) and single-cell chromatin immunoprecipitation sequencing (scChIP-seq) (Rotem et al., 2015) which can measure the H3 lysine 4 tri-methylation (H3K4me3) and di-methylation (H3K4me2) modifications. For more detail, please see the Epigenome sequencing section.

Based on these single-cell epigenomic technologies, several single-cell multi-modal high throughput methods targeting both chromatin accessibility and transcriptome have been developed. Cao et al. developed sci-CAR (Cao et al., 2018), the first protocol that can jointly profile the mRNA and ATAC in the same cell. SciCAR applied combinatorial indexing for each cell, and then redistributed cells by FACS and lysate splitting, and then amplified for sequencing. Cao et al. applied sci-CAR to human and mouse cell line mixture and mouse kidney tissue and identified cis-regulatory network from the joint profiling datasets. However, due to the high sparsity in the scATAC modality and limited coverage in scRNA modality of sci-CAR, only a minority of differentially accessible sites and differentially expressed genes in bulk scRNA and scATAC sequencing can be discovered by single-cell datasets in sci-CAR (Cao et al., 2018). Chen et al. developed the droplet-based

4295 single-nucleus chromatin accessibility and mRNA expression sequencing
 4296 (SNARE-seq) (Chen et al., 2019d), which enhanced the sequencing coverage of both
 4297 scRNA and scATAC, and improved the coverage limitation in sci-CAR. SNARE-seq
 4298 used a splint oligonucleotide with sequence complementary to the adaptor sequence
 4299 inserted by ATAC transposition (5' end) and the mRNA poly(A) bases (3' end), which
 4300 allowed to capture both omics data. Compared to sci-CAR, SNARE-seq detected 4-5
 4301 times more chromatin accessible sites in the mouse postnatal brain dataset and adult
 4302 brain dataset than sciCAR tissue dataset, and improved the throughput by a cellular
 4303 combinatorial indexing strategy (Preissl et al., 2018). Zhu et al. further improved the
 4304 protocol and developed the parallel analysis of individual cells for RNA expression
 4305 and DNA accessibility by sequencing (Paired-seq), which adopted a ligation-based
 4306 combinatorial indexing strategy to simultaneously tag both the open chromatin
 4307 fragments by the Tn5 transposases and the cDNA molecules by reverse transcription
 4308 (RT) of RNA. Zhu et al. applied Paired-seq to mouse embryonic cerebral cortex tissue
 4309 and applied integrated analysis with ENCODE mouse embryonic cerebral cortex
 4310 tissue datasets, reconstructed the cellular trajectory, and recovered the cis-regulatory
 4311 network from the dual-omic dataset (Zhu et al., 2019b). More recently, Ma et al.
 4312 developed simultaneous high-throughput ATAC and RNA expression with sequencing
 4313 (SHARE-seq) (Ma et al., 2020), which used multiple rounds of hybridization blocking
 4314 to joint labeling mRNA and chromatin fragments in the same single cell. Compared to
 4315 sciCAR (Cao et al., 2018), SNARE-seq (Preissl et al., 2018), and Paired-seq (Zhu et
 4316 al., 2019b), SHARE-seq showed higher scalability on much larger library size for
 4317 more than 30,000 cells and higher sensitivity with more genes and ATAC peaks
 4318 detected in each cell than previous multi-modal methods. Based on higher data quality,
 4319 Ma et al. applied a new definition of domains of regulatory chromatin (DORCs) rather
 4320 than individual peaks to analyze the regulatory map between chromatin accessibility
 4321 and gene expression, and identified prior functions of DORCs to gene expression in
 4322 cell lineage choice and cell fate decision (Preissl et al., 2018). Recently, 10x genomics
 4323 developed 10x Multiome, a commercial service platform for joint profiling of scRNA
 4324 and scATAC in the single cell, which would accelerate the applications of scRNA and
 4325 scATAC multi-modal techniques in more biological and clinical research.

4326

4327 *Transcriptome + proteome*

4328 Besides single-cell multi-modal technologies targeting DNA and RNA molecules,
 4329 several single-cell multi-modal methods that can measure RNA and protein
 4330 simultaneously in the same cell were developed. Peterson et al. developed proximity
 4331 extension assay/specific RNA target amplification (PEA/STA) (Genshaft et al., 2016)
 4332 method. PEA/STA applied reverse transcriptase as the DNA polymerase for both RT
 4333 of RNA and extension of proximity extension assay (PEA) DNA oligos for 38
 4334 proteins to enable cDNA synthesis and PEA to proceed in the same reaction in the
 4335 Fluidigm C1™ system (DeLaughter, 2018). Andreas et al. developed proximity
 4336 ligation assay for RNA (PLAYR) (Frei et al., 2016), a method for highly multiplexed
 4337 transcript quantification using flow and mass cytometry, which is also compatible
 4338 with standard antibody staining. Using the mass cytometry, PLAYR allowed

4339 simultaneous measurement of more than 40 mRNAs and proteins, and enabled the
 4340 characterization of the interplay between transcription and translation at single-cell
 4341 level. Besides protein-DNA ligation strategy, two methods targeting surface protein
 4342 and mRNA, cellular indexing of transcriptomes and epitopes by sequencing
 4343 (CITE-seq) (Stoeckius et al., 2017), and RNA expression and protein sequencing
 4344 assay (REAP-seq) (Peterson et al., 2017) were developed to detect both mRNAs and
 4345 cell surface proteins using oligonucleotide-labeled antibodies, enabled the multimodal
 4346 analysis at single-cell scale by droplet-based single-cell sequencing approaches. These
 4347 two methods greatly improved the throughput of the transcriptome. For example,
 4348 REAP-seq can quantify proteins with 82 barcoded antibodies and measure more than
 4349 20,000 genes in a single workflow (Peterson et al., 2017). Another method, single-cell
 4350 RNA and immunodetection (RAID) (Gerlach et al., 2019), can detect intracellular
 4351 proteins and phosphorylated proteins together with mRNAs. RAID immunostained
 4352 the intracellular target proteins with antibodies conjugated with RNA barcodes, and
 4353 then converted proteins into RNAs.

4354

4355 *Techniques capturing more than two modalities*

4356 Based on these bi-modal single-cell methods we discussed above, several methods
 4357 were developed to capture more than two omics in the same cell. Single-cell
 4358 nucleosome, methylation and transcription sequencing (scNMT-seq) (Clark et al.,
 4359 2018) were developed by combining scM&T-seq (Angermueller et al., 2016) and
 4360 NOME-seq to measure nucleosome, transcriptome, and DNA methylome in the same
 4361 cell. Recently, Wang et al. developed scNOMeRe-seq (Wang et al., 2021b), which
 4362 enabled the profiling of genome-wide chromatin accessibility, DNA methylation, and
 4363 transcriptome in the same individual cell and applied this method for a single-cell
 4364 multi-omics map of mouse preimplantation development. Based on CITE-seq,
 4365 Mimitou et al. developed expanded CRISPR-compatible cellular indexing of
 4366 transcriptomes and epitopes by sequencing (ECCITE-seq) to capture mRNA, sgRNA
 4367 and designed target proteins for at least five modalities in the same single cell. By
 4368 adapting Paired-seq (Zhu et al., 2019b) with cleavage under targets and tagmentation
 4369 (CUT&Tag) strategy (Kaya-Okur et al., 2019), Zhu et al. further developed parallel
 4370 analysis of individual cells for RNA expression and DNA from targeted tagmentation
 4371 by sequencing (Paired-Tag), (Zhu et al., 2021a) a novel protocol which can
 4372 simultaneously profile scRNA, scATAC and five histone modifications in the same
 4373 cell. Also, Zhang et al. developed scCUT&Tag-pro (Zhang et al., 2022a), which
 4374 combined CUT&Tag with CITE-seq and captured five histone modifications by
 4375 CUT&Tag library and proteins by antibody-derived protein tags library.

4376

4377 *Multi-omics integration analysis*

4378 Recent advances in single-cell multi-modal technologies provided substantial data
 4379 resources to uncover the molecular mechanism by multi-view high-dimensional and
 4380 high-resolution datasets. However, it is hard to properly integrate the multi-modal
 4381 single-cell datasets, arising from high dataset dimensionality, high data sparsity as
 4382 well as complex variables among multimodal datasets and techniques. An increasing

number of algorithms were developed for different applications and tasks in multi-modal data integration analysis, indicating the rapid progression as well as growing attention from researchers to the field of single-cell multi-modal data integration and data analysis. Also, several review papers (Adossa et al., 2021; Argelaguet et al., 2021; Lance et al., 2022; Xu and McCord, 2022) and benchmark studies (Brombacher et al., 2022; Luecken et al., 2022) have performed specialized classification as well as general evaluations for emerging multi-modal integration tools, providing great benefits for other researchers to select appropriate methods from different types of integration tools. Here, we introduce a comprehensive set of multimodal integration tools and related studies from the perspectives of the following two sections, including 1) categories of multimodal integration tools, which introduce recently published single-cell multi-modal integration tools by different classification standards, and 2) recent benchmark studies for these recently published tools.

4397

4398 *Categories of multimodal integration tools*

4399 Based on previous review papers (Adossa et al., 2021; Argelaguet et al., 2021; 4400 Stanojevic et al., 2022), several criterions can be applied to classify multimodal 4401 integration tools. Firstly, multimodal integration tools can be classified with four 4402 major integration strategies based on the choice of the shared features (known as 4403 anchors) for data integration, including horizontal integration with all genomic 4404 features like genes as anchor, vertical integration with all common cells as anchor, 4405 diagonal integration with no shared features and mosaic integration with partially 4406 shared cells and partially shared genomic features as anchor (Argelaguet et al., 2021). 4407 Several tools with similar methodologies can also be developed as different types of 4408 integration tools. For example, for algorithms based on non-negative matrix 4409 factorization (NMF), iNMF (Gao et al., 2021a) was designed as vertical integration 4410 tool, coupledNMF (Gao et al., 2021a) was designed as diagonal integration tool, and 4411 UINMF (Kriebel and Welch, 2022) was designed as mosaic integration tool.

4412 Secondly, the multimodal integration tools can also be grouped by the types of 4413 multi-modal techniques, including “paired” and “unpaired” integration tools 4414 (Brombacher et al., 2022; Stanojevic et al., 2022). The paired multimodal integration 4415 tools were specifically designed for multimodal datasets simultaneously captured and 4416 sequenced from the same cell. The vertical integration and mosaic integration 4417 strategies are usually applied in paired integration tools, as the paired datasets share 4418 all or partial common cells between different modalities. The unpaired integration 4419 tools were designed for integrating independent single-cell experiments from different 4420 modalities, as the cells of one modality cannot find matched cells from the other 4421 modalities. Due to the difference in both cells and features, diagonal integration was 4422 commonly applied for unpaired data integration. Several tools were developed for 4423 paired or unpaired multimodal dataset integration specifically. For example, the 4424 Seurat v3 (Stuart et al., 2019) was designed for unpaired multi-modal datasets, and the 4425 updated version – Seurat v4 (Hao et al., 2021b) was specifically designed for paired 4426 multi-modal datasets. Also, some integration tools (Zhang et al., 2021c; Hu et al.,

4427 2022b; Lin et al., 2022) can be applied to both paired and unpaired multi-modal
 4428 datasets.

4429 Thirdly, based on the methodologies, the multimodal integration tools can be
 4430 categorized into several sub-types (Stanojevic et al., 2022), including mathematical
 4431 matrix factorization methods, manifold alignment methods, network-based methods,
 4432 and deep learning methods. The deep learning integration tools can be further
 4433 categorized by the infrastructure of the deep model, including autoencoder (AE),
 4434 generative adversarial network (GAN), graph neural network (GNN), and their
 4435 extended structure such as variational autoencoder (VAE). The selection of the
 4436 methodology is largely determined by the type of multimodal dataset and the
 4437 integration tasks. For unpaired multimodal datasets, manifold alignment methods can
 4438 first reduce the different features of multimodal datasets into the same dimension of
 4439 latent embeddings/manifolds, and then integrate the heterogeneous modalities by the
 4440 same manifolds (Argelaguet et al., 2021; Stanojevic et al., 2022). Similarly, matrix
 4441 factorization methods can be applied for different integration tasks by matrix
 4442 factorizing unmatched features or cells to the matrix of the same dimension with less
 4443 information loss than simple dimension reduction methods along with manifold
 4444 alignment (Stanojevic et al., 2022). The deep learning tools using GAN (Xu et al.,
 4445 2021b; Khan et al., 2022; Zhao et al., 2022a), VAE (Ashuach et al., 2021; Minoura et
 4446 al., 2021; Lotfollahi et al., 2022) and transformer (Li et al., 2022b) can learn the
 4447 common latent embedding from different modalities of same cells or shared cells and
 4448 then impute the missing cells and features, as GNN (Ma et al., 2021a; Cao and Gao,
 4449 2022) model is applied to learn the relationship between different types of features
 4450 (gene in scRNA and peak in scATAC for etc.) and infer biological network in
 4451 multimodal data integration (Ma et al., 2021a; Cao and Gao, 2022).

4452 Fourthly, the multimodal integration tools can be classified based on certain
 4453 omics for integration. Several multimodal integration algorithms were designed for
 4454 specific multi-modal datasets integration, for example, CiteFuse (Kim et al., 2020a)
 4455 was designed for CITE-seq (Peterson et al., 2017) analysis, SCIM (Stark et al., 2020)
 4456 was designed for scRNA and CyTOF integration, while scMVP (Li et al., 2022b) was
 4457 designed for paired scRNA and scATAC datasets integration. Also, besides these tools
 4458 restricted to specific omics datatypes, several algorithms like LIGER (Welch et al.,
 4459 2019) were designed for general integration tasks without restrictions on integration
 4460 omics types.

4461 The multimodal integration tools can be also categorized by major coding
 4462 languages like Python and R, and special integration applications like cross modality
 4463 translation. All multimodal integration tools along with their categories are
 4464 summarized in Table S10.

4465

4466 *Benchmarks for single-cell multimodal integration tools*

4467 Although plenty of multimodal integration studies have been published for different
 4468 tasks in multi-modal single-cell analysis, it is still difficult to find state-of-the-art
 4469 methods from published integration methods. To solve this issue, recently, several
 4470 benchmark studies have been performed for the evaluation of single-cell multimodal

4471 integration tools for different tasks of multi-modal dataset analysis (Brombacher et al.,
 4472 2022; Lance et al., 2022; Luecken et al., 2022). These benchmark studies would
 4473 provide comprehensive and objective evaluations of the performance of these
 4474 candidate integration tools from the perspective of data users. Next, we introduce
 4475 recent benchmark studies and summarize the performance of multimodal integration
 4476 tools from these third-party evaluation studies.

4477 Malte et al. performed a benchmark analysis of single-cell integration tools for
 4478 tasks of atlas level data integration and developed a benchmark pipeline for objective,
 4479 comprehensive, and reproducible evaluation of single-cell integration tools (Luecken
 4480 et al., 2022). This study included several unpaired multimodal single-cell integration
 4481 tools, however, it only focused on the integration tasks for different datasets from the
 4482 same modalities, like integration of different scRNA datasets and integration of
 4483 different scATAC datasets, but did not provide evaluation for cross modality
 4484 integration of paired or unpaired scRNA and scATAC data. Nevertheless, this study
 4485 provided a stable, comprehensive, and highly scalable **benchmark** framework for
 4486 single-cell atlas integration evaluation.

4487 Recently, to better accomplish the analysis challenges arising from data sparsity,
 4488 technical and biological variability, and high dimensionality from single-cell
 4489 multimodal integration analysis, NeurIPS2021 launched an online competition for
 4490 three major tasks in single-cell multimodal data integration analysis (Lance et al.,
 4491 2022), including 1) predicting one modality from another, 2) matching cells between
 4492 modalities, and 3) jointly learning representations of cellular identity. Among the
 4493 three tasks, the second task is specifically designed for unpaired multimodal
 4494 integration tools, and the third task is designed for paired multimodal integration tools.
 4495 Also, the competition launcher generated the first single-cell multimodal
 4496 benchmarking datasets, including a multi-center CITE-seq dataset with 90,000 cells
 4497 for scRNA and protein integration tasks and a multi-center 10x Multiome dataset with
 4498 70,000 cells for scRNA and scATAC integration tasks. Among all three tasks, the
 4499 CLUE (Cross-Linked Universal Embedding) algorithm, a semi-supervised modality
 4500 matching function in GLUE (Cao and Gao, 2022) package, won the first prize and got
 4501 the all categories winner in the second modality aligning task, showing the top
 4502 performance of cross modality matching among unpaired integration tools (Lance et
 4503 al., 2022). However, as the competition only evaluated the algorithms from online
 4504 submitters, most published single-cell multimodal integration tools were not included
 4505 in the benchmark evaluation.

4506 For further comparison of published single-cell multimodal integration tools with
 4507 deep learning framework, Eva et al. first reviewed 18 recently published multimodal
 4508 integration tools using deep learning model, and then performed the first
 4509 comprehensive benchmark study for selected popular tools using CITE-seq dataset
 4510 and 10x Multiome dataset from NeurIPS2021. For biology preservation tasks, Cobolt
 4511 (Gong et al., 2021) showed the best performance among benchmark algorithms on
 4512 both CITE-seq and 10x Multiome datasets, only for larger cell numbers, scMVP (Li et
 4513 al., 2022b) has better performance than Cobolt (Gong et al., 2021) on 10x Multiome
 4514 dataset. For technique effect removal tasks, SCALEX (Xiong et al., 2021b) showed

4515 consistent top performance on CITE-seq dataset, and scMVP (Li et al., 2022b)
4516 showed the highest performance on 10x Multiome dataset.

4517

4518 *Summary*

4519 In this chapter, we have provided a summary of recent advances in multiple types of
4520 multi-omic single-cell sequencing techniques, and their bioinformatics integration
4521 methods. With improvements in experimental data quality and the performance of
4522 bioinformatics algorithms, single-cell multi-modal technologies have provided
4523 comprehensive multi-modal insights into different eras at the single-cell level.
4524 Moving forward, expanding the biological applications of single-cell multimodal
4525 techniques, as well as increasing the performance and robustness of single-cell
4526 multi-modal algorithms, would undoubtedly accelerate new discoveries in biological
4527 and medical research. These improvements hold significant potential to revolutionize
4528 our understanding of cellular processes and the development of personalized
4529 medicine.

4530

4531 **Chapter 7 Single-cell spatial transcriptomics technology**

4532

4533 In the preceding sections, we systematically reviewed the recent advances in single-cell
4534 omics. Although these single-cell sequencing technologies allow the investigation of
4535 cellular heterogeneity at an unprecedented resolution, they are far from adequate to get
4536 a full understanding of the intricate workings of multicellular organisms. Many studies
4537 emphasize that the state of one cell is not only regulated by the intracellular regulatory
4538 network but also interfered by the extracellular signals from the environment (Lin and
4539 Hankenson, 2011; Juntila and de Sauvage, 2013; Dries et al., 2021a). Both the
4540 dissociation of tissues and the isolation of single cells during experimental procedures
4541 cause the loss of critical spatial information, including cell positions and their mutual
4542 proximities. Spatial transcriptomics (ST) has addressed this limitation, enabling the
4543 measurement of gene expression with spatial information preserved. In this section, we
4544 will introduce spatial transcriptomics technologies, discuss computational methods for
4545 spatial data analysis, and provide a review of their applications in various biological
4546 systems. Additionally, we will also delve into the current progress in techniques of
4547 spatial multi-omics.

4548

4549 *Techniques for spatially resolved transcriptomics*

4550 All current spatial transcriptomics techniques can be broadly summarized into three
4551 categories majorly based on 1) microdissection, 2) barcoding, and 3) imaging,
4552 respectively (Figure 14A-C). These ST technologies differ in their approaches to
4553 location labeling and transcript profiling, which may determine the spatial resolution,
4554 detection efficiency, demanding sample types, and so on. Next, we will discuss the
4555 principles of a selection of representative techniques from each category and
4556 summarize their characteristics. The curated list of techniques and their corresponding
4557 features is shown in Table 4.

4558

4559 *Microdissection-based ST techniques*

4560 Techniques falling within this category aim to computationally reconstruct the 3D
 4561 structure of tissues from multiple spatially proximal tissue subregions isolated by
 4562 various microdissection approaches (Figure 14A). For instance, RNA tomography
 4563 (tomo-seq) obtains RNA from a series of sequential cryosections along three
 4564 orthogonal axes in multiple putatively identical biological samples (Junker et al., 2014).
 4565 The requirement of identical biological samples limits the application of tomo-seq on
 4566 human samples. By comparison, STRP-seq slices tissues into primary sections and then
 4567 secondary stripes using a two-level dissection strategy, which assumes that spatial
 4568 expression patterns are constant between consecutive primary sections spaced 14 μm
 4569 apart (Schede et al., 2021). Based on cryosectioning, Geo-seq utilizes laser capture
 4570 microdissection (LCM) to section tissues into regions as small as around 10 cells (Chen
 4571 et al., 2017b). Other methods within this type includes ProximID (Boisset et al., 2018)
 4572 and PIC-seq (Giladi et al., 2020), which focus on physical cell interaction within two
 4573 (doublets) or three cells (triplets), rather than the positions or surrounding context in the
 4574 tissue.

4575 In addition to physical sectioning, microdissection could be accomplished by
 4576 combining optical marking with fluorescence-based cell selection, or photo-cleavage of
 4577 gene index oligos. For example, transcriptome *in vivo* analysis (TIVA) loads TIVA tags
 4578 (i.e., photoactivatable mRNA capture molecules) into live cells and selects cells by
 4579 laser photoactivation, which subsequently triggers tags' hybridization to mRNA
 4580 (Lovatt et al., 2014). As an alternative technology, NICHE-seq injects labeled landmark
 4581 cells into transgenic mice expressing photoactivatable green fluorescent (PA-GFP),
 4582 allowing *in situ* labeling of niches of interest. After tissue dissociation, activated
 4583 PA-GFP⁺ cells are sorted by fluorescence-activated cell sorting (FACS) for single-cell
 4584 transcriptome profiling (Medaglia et al., 2017). The commercial GeoMX Digital
 4585 Spatial Profiler (DSP) developed by nanoString employs probes with UV cleavable
 4586 linkers and automates the optical selection (Merritt et al., 2020).

4587 Overall, microdissection coupled with single-cell or bulk RNA sequencing makes
 4588 it possible to study the transcriptome within the spatial context. Microdissection can be
 4589 performed in a physical manner, or an optics-dependent way. The physical sectioning is
 4590 often implemented manually, making the dissection protocol labor-intensive and
 4591 time-consuming. In contrast, optics-dependent sectioning generally depends on the
 4592 loading of specialized tags into live cells or genetic engineering in model organisms,
 4593 which restricts its application to fresh-frozen or FFPE human samples. No matter how
 4594 microdissection and sequencing are performed, the exact positions of profiled cells
 4595 within the selected subregion are unknown, resulting in a generally low spatial
 4596 resolution.

4597

4598 *Barcoding-based ST techniques*

4599 Microdissection-based techniques trace the spatial information by manually labeling
 4600 each subregion. The spatial barcoding techniques enable automatic recording of spatial
 4601 coordinates (Figure 14B). In 2016, Ståhl et al. pioneered the application of barcoding
 4602 techniques in the ST technology (Ståhl et al., 2016). In this approach, barcodes,

4603 together with unique molecule identifiers (UMIs) and ploy-T oligonucleotides are
 4604 immobilized on glass slides to allow *in situ* capture of mRNA and cDNA synthesis.
 4605 Each barcoded spot in the array is 100 μm in diameter and is positioned 200 μm
 4606 center-to-center apart from the adjacent ones, which brings about a resolution of 10~40
 4607 cells. 10x Genomics has further enhanced the spatial resolution to 5~10 cells using
 4608 spots with diameters of 55 μm and center-to-center distances of 100 μm . Instead of
 4609 directly attaching barcodes to slides, some techniques link barcodes to beads for
 4610 position labeling and mRNA capture. For example, Slide-seq deposits 10- μm
 4611 DNA-barcoded beads onto a surface (Rodriques et al., 2019; Stickels et al., 2021).
 4612 Similarly, HDST places barcoded beads into an array with 2- μm wells (Vickovic et al.,
 4613 2019). Both of these technologies improve spatial resolution to 1~2 cells. However, as
 4614 the barcoded beads are randomly distributed on the slide, *in situ*, sequencing (ISS) or *in*
 4615 *situ* hybridization (ISH) is required to decode each fixed bead's barcode sequence.

4616 Although bead-based techniques can reach a cellular resolution, they are still too
 4617 coarse to detect subcellular differences. Recently, Seq-scope was developed to achieve
 4618 a 0.5~0.8 μm center-to-center resolution by repurposing the Illumina sequencing
 4619 platform (Cho et al., 2021). Another technique enabling sub-micrometer-resolution
 4620 profiling is Stereo-seq (Chen et al., 2022), where 220 nm DNA nanoballs (DNBs)
 4621 containing barcodes are deposited on a patterned array with a center-to-center distance
 4622 of 500 or 715 nm. Both Seq-scope and Stereo-seq require two-round sequencing, in
 4623 which the first associated barcodes with spatial locations and the second provides
 4624 information of captured cDNA, as performed in Slide-seq.

4625 To summarize, the barcoding-based approaches combine spatial barcoding
 4626 techniques with NGS to allow transcriptome-wide profiling of RNA in the spatial
 4627 context. The technologies involve a trade-off between spatial resolution and detection
 4628 efficiency. Compared to the original ST technology or commercialized 10x Visium, the
 4629 improvement in spatial resolution by Seq-scope, Stereo-seq tends to come at the
 4630 expense of low detection sensitivity and low gene coverage.

4631

4632 *Imaging-based ST techniques*

4633 Both microdissection-based and barcoding-based techniques extract nucleic acid
 4634 molecules for NGS sequencing after position labeling. To preserve RNA *in situ*, various
 4635 *in situ* transcriptomic techniques were developed for the spatial mapping of gene
 4636 expression, including ISH and ISS (Figure 14C). Because these methods necessitate
 4637 fluorescence imaging, they are collectively referred to as imaging-based techniques.

4638 Most ISH-based ST techniques mainly rely on single-molecule RNA fluorescence
 4639 *in situ* hybridization (smFISH) (Femino et al., 1998) to enable quantitative
 4640 measurements of targeted transcripts *in situ*. SeqFISH belongs to this type, which
 4641 allows the simultaneous detection of multiple mRNA molecules by sequential rounds
 4642 of fluorescent hybridization, imaging, and stripping of readout probes (Lubeck et al.,
 4643 2014). Using the seqFISH strategy, all the targeted genes are encoded by the
 4644 combination of rounds of readout probes. SeqFISH+ expands the readout probe palette
 4645 from four or five colors in seqFISH to 60 ‘pseudo colors’ (Eng et al., 2019), enabling
 4646 the multiplexing of up to 10,000 genes in a single cell. MERFISH is another

4647 smFISH-based technique, which also requires multiple rounds of hybridization, but
 4648 employs a distinct multi-bit binary encoding strategy (Chen et al., 2015a). To address
 4649 the issue of optical crowding, expansion microscopy (ExM) was integrated into
 4650 MERFISH (Xia et al., 2019). The encoding strategy, in conjunction with ExM, allows
 4651 MERFISH to reduce the number of hybridization rounds. For example, to ensure the
 4652 detection of ~10,000 genes, using 3-color imaging, seqFISH+ needs 80 (4*20) rounds
 4653 of hybridization, while MERFISH only needs 23 rounds to construct a 69-bit HD4 code
 4654 with a Hamming weight of 4 (Zhuang, 2021).

4655 In addition to the techniques based on multiplexed FISH, *in situ* profiling of RNA
 4656 can also be achieved by ISS, which sequences RNA in the fixed tissue or cell sample
 4657 with *in situ* signal amplification. Due to the limited cellular space, some of the
 4658 ISS-based techniques select a part of genes by designing probes to target specific RNA
 4659 or cDNA. The initial ISS approach published in 2013 uses padlock probes to bind to
 4660 targets (Ke et al., 2013), followed by rolling-circle amplification (RCA) to generate
 4661 RCA products for subsequent sequencing by ligation. STARmap uses two-component
 4662 padlock probes to directly bind to RNA rather than cDNA, avoiding the inefficient step
 4663 of RNA-to-cDNA and reducing potential noise (Wang et al., 2018). To diminish strong
 4664 background fluorescence brought by conventional supported oligo ligation detection
 4665 (SOLiD) sequencing, sequencing with error-reduction by dynamic annealing and
 4666 ligation (SEDAL) was devised for STARmap, which allows error rejection during
 4667 sequencing.

4668 Besides targeted ISS methods, ISS could be conducted in an untargeted manner, in
 4669 which transcripts are reversely transcribed to cDNA, followed by DNA amplification
 4670 and sequencing without pre-selection of genes. While the untargeted manner could
 4671 improve the coverage to transcriptome-wide, it can also lead to molecule crowding. To
 4672 mitigate it, FISSEQ leverages a partition sequencing strategy (Lee et al., 2015), where
 4673 only a small fraction of amplicons is randomly selected and sequenced using extended
 4674 sequencing primers, therefore resulting in low detection efficiency. Combined with
 4675 ExM, FISSEQ was adapted to another approach called ExSeq to discriminate between
 4676 crowded molecules and increase spatial resolution (Alon et al., 2021).

4677 In general, imaging-based techniques offer high spatial resolution, reaching
 4678 cellular or even subcellular levels. Among these techniques, ISH-based ones, which
 4679 rely on prior knowledge of target genes, exhibit high detection efficiency. By
 4680 comparison, due to the limitations of ISS, ISS-based techniques have comparatively
 4681 low efficiency, especially in untargeted ones. Moreover, most of these techniques
 4682 necessitate specialized equipment for high-resolution imaging, which may limit their
 4683 broader applicability.

4684

4685 *Techniques for spatial multi-omics*

4686 To achieve a more comprehensive characterization of cells, considerable efforts have
 4687 been directed towards the measurement of other modalities in the spatial context,
 4688 including genome, epigenome, proteome, metabolome, and so on (Figure 14D). The
 4689 positioning strategies used in ST technologies have been adapted to realize spatial
 4690 profiling of other omics. For instance, Slide-DNA-seq captures spatially resolved

4691 genomic sequences using a barcoded bead array which was initially developed for
 4692 spatial RNA profiling (Zhao et al., 2022b). Similarly, spatial-CUT&Tag (Deng et al.,
 4693 2022a) and spatial-ATAC-seq (Deng et al., 2022b) were developed to profile histone
 4694 modification and chromatin accessibility by combining DbiT-seq's microfluidic
 4695 deterministic barcoding strategy (Liu et al., 2020b) with *in situ* CUT&Tag chemistry
 4696 and Tn5 transposition chemistry, respectively. To gain an understanding of 3D
 4697 chromatin conformation within its native context, a MERFISH-based method was
 4698 designed to visualize over 1,000 genomic loci for high-resolution chromatin tracing (Su
 4699 et al., 2020).

4700 In the realm of proteome, protein expression could be readily visualized by
 4701 multiplexed immunohistochemistry (IHC). IHC can be further coupled with imaging
 4702 mass cytometry (Giesen et al., 2014) or multiplexed ion beam imaging (MIBI) (Angelo
 4703 et al., 2014) to allow the simultaneous imaging of ~100 proteins. Moreover, proteins of
 4704 interest can be targeted by DNA-barcoded antibodies and thus quantified by NGS, as in
 4705 GeoMx DSP (Merritt et al., 2020). Cell-surface proteins can be bound by antibodies
 4706 without cell lysis, preventing damage to RNA. Therefore, proteomics could be
 4707 combined with transcriptomics in both single-cell and spatial omics. For example, the
 4708 enhanced version of 10x Visium conducts IHC prior to mRNA capturing to enable the
 4709 co-detection of protein and RNA, albeit only allowing for the detection of 1~2 proteins.
 4710 By adding antibody-derived tags to fixed tissue slides before flow barcoding, DbiT-seq
 4711 enables the co-measurement of mRNA and dozens of proteins (Liu et al., 2020b).
 4712 Additionally, nanoString offers the CosMx SMI platform, enabling the quantification
 4713 of 1,000 RNA and 64 protein analytes through high-plex imaging (He et al., 2021a).

4714 Metabolites collected from samples are often quantified using mass spectrometry
 4715 (MS). For the study of spatially resolved metabolome, various techniques have been
 4716 developed based on imaging mass spectrometry (IMS). These techniques differ in the
 4717 manners by which ions are produced from molecules of samples, including
 4718 matrix-assisted laser desorption/ionization (MALDI) (Rappez et al., 2021), desorption
 4719 electrospray ionization (DESI) (Yin et al., 2019b), and secondary ion mass
 4720 spectrometry (SIMS) (Passarelli et al., 2017). For example, SpaceM is a MALDI-based
 4721 method for *in situ* single-cell metabolomics (Rappez et al., 2021). It addresses the
 4722 challenge of assigning metabolite intensities to individual cells by integrating
 4723 MALDI-imaging with light microscopy followed by computational methods for image
 4724 segmentation and registration.

4725 In addition to intrinsic genetics, many gene functions are influenced by the spatial
 4726 context (Haigis et al., 2019). To study spatial functional genomics, Dhainaut et al.
 4727 established an approach called Perturb-map (Dhainaut et al., 2022), which enables
 4728 pooled CRISPR screens at the single-cell resolution in the tissue context. This is
 4729 achieved by employing a protein barcode system and multiplex imaging.
 4730

4731 *Computational methods for spatial transcriptomics*

4732 A standard workflow for single-cell analysis encompasses critical tasks such as cell
 4733 clustering, cell-type annotation, differential expression analysis, lineage tracing,
 4734 cell-cell communication, and integration analysis. These tasks also form the backbone

of ST data analysis. Spatial transcriptomics, with its unique capacity to provide information about spatial proximity and context, not only broadens the analytical scope but also poses great integration challenges. To address these, a large number of computational methods have been developed to integrate gene expression with spatial information and provide new insights (Figure 15). We will review the methods designed for different purposes in the forthcoming sections. A list of published computational methods is presented in Table S11.

4742

4743 *Denoising to enhance the signal in spatial transcriptomics*

4744 As discussed above, many ST techniques face challenges related to low detection
 4745 efficiency and significant noise. These issues arise from shallow sequencing for each
 4746 spatial unit (i.e., spot or bead) or complex experimental steps needed to preserve the
 4747 tissue structure, or a combination of both (Wang et al., 2022f). Wang et al. have shown
 4748 in 10x Visium and Slide-seq data that the signal noise was reflected in both the
 4749 drop-outs and the inflation of gene expression (Wang et al., 2022f). While denoising
 4750 methods have been developed for scRNA-seq data to address the drop-out problem,
 4751 they often struggle to correct the ‘false’ high expression. Furthermore, these single-cell
 4752 methods rely solely on transcriptomics data, and thus could not be directly applied to
 4753 integrate additional spatial information.

4754 Several computational methods have been specially developed to tackle the
 4755 denoising ST data. For example, Sprod could impute gene expression in noisy ST data
 4756 from barcoding-based techniques based on latent graph learning (Wang et al., 2022f).
 4757 The denoising process in Sprod involves two steps. First, Sprod builds a graph by
 4758 leveraging spatial proximity and expression similarity. Importantly, if available,
 4759 features extracted from the corresponding pathological images could be incorporated
 4760 for graph construction. Next, Sprod corrects gene expression for each spot/bead by
 4761 borrowing expression information from the neighborhood revealed in the graph.
 4762 Another method, spARC, adopts a similar graph-based framework but demonstrates its
 4763 serviceability on imaging-based ST techniques (Kuchroo et al., 2022). SiGra is also a
 4764 graph-based method but employs a different approach to build the graph (Tang et al.,
 4765 2022). It utilizes three graph transformer autoencoders for imaging, transcriptomics,
 4766 and hybrid, respectively, as well as the attention mechanism, enabling SiGra to enhance
 4767 the sparse and noisy transcriptomics data with multi-modal spatial information. The
 4768 SME method in stLearn also allows the integration of image features to normalize
 4769 spatial gene expression (Pham et al., 2020). It employs a simple strategy of weighted
 4770 average, where the weights are calculated based on the morphological similarity
 4771 between close spots.

4772 Rather than random drop-outs or inflation, Ni et al. believed that the loss and
 4773 inflation are caused by the bleed of mRNA between and among nearby spots, which is
 4774 referred to as spot swapping (Ni et al., 2022). To adjust for the effects of spot swapping,
 4775 they proposed a method called SpotClean (Ni et al., 2022). SpotClean employs a
 4776 probabilistic framework to model gene-specific expression at a given spot, which
 4777 considers reads present in tissue at that spot and reads bleeding out into other spots, and
 4778 also removes reads bleeding in from other spots. The authors demonstrated that

4779 SpotClean could accurately estimate gene-specific UMI counts in technologies such as
 4780 ST and 10x Visium, where background positions could be identified by the alignment
 4781 between the ST slide and the matched H&E image.

4782

4783 *Subcellular analysis for imaging-based ST data*

4784 Imaging-based ST techniques provide good opportunities for cellular or even
 4785 subcellular analyses but also pose great challenges. For these technologies, every
 4786 measured pixel represents only one transcript, which is insufficient to infer the cell type
 4787 it belongs to. How to merge these single pixels to form a cell or a sub-cellular structure
 4788 will be of great significance. In the current studies, two primary strategies are utilized
 4789 for the analysis of high-resolution ST data: segmentation-based or segmentation-free
 4790 approaches.

4791

4792 *Cell segmentation-based analysis*

4793 Cell segmentation, initially proposed in the processing of microscopic IHC images,
 4794 provides more information about cell number and cell morphology. Cell segmentation
 4795 here is to determine the cell boundaries based on the sparse measurement of transcripts,
 4796 namely to assign transcripts to cells. Conventional cell segmentation relies on features
 4797 extracted from the staining images, including intensity and textures, some of which
 4798 could represent the cell boundaries. But for the fluorescent images of RNA, revealing
 4799 cell boundaries requires specific staining for cell membranes, which hampers the
 4800 segmentation of cells. Most groups choose to perform additional nucleus staining (e.g.,
 4801 DAPI) to identify putative nuclei, which are then used to guide the transcript
 4802 assignment (Wang et al., 2018; Eng et al., 2019). Considering that the gene expression
 4803 in the nucleus region may not equate to the expression within the whole cell, some of
 4804 the groups combine auxiliary ploy(A) staining to inform the soma of cells (Moffitt et
 4805 al., 2018; Wang et al., 2018). Several computational methods have been developed to
 4806 provide alternative solutions.

4807 For instance, Qian et al. developed pciSeq, which utilized a probabilistic
 4808 framework to assign RNA spots to their original cells (Qian et al., 2020). To be specific,
 4809 pciSeq treats the nuclei segmentation from the DAPI image as the initial approximation
 4810 of cells, and models the cellular RNA counts and the gene-cell distance by a negative
 4811 binomial distribution and a Poisson process, respectively. With paired scRNA-seq as a
 4812 reference, the method estimates the probability of a transcript belonging to a cell and a
 4813 cell type simultaneously using variational Bayes inference. JSTA is another method
 4814 relying on initial nucleus segmentation from DAPI staining and matched scRNA-seq
 4815 reference (Littman et al., 2021). JSTA can also achieve joint cell segmentation and cell
 4816 type annotation through iterative pixel assignment with a deep neural network (DNN)
 4817 as a classifier.

4818 Cell segmentation can be implemented in a scRNA-independent way. For example,
 4819 Baysor can perform cell segmentation based on the expression of transcripts alone
 4820 (Petukhov et al., 2022), while also supporting the integration with prior information
 4821 from cell-type-specific expression profiles obtained by scRNA-seq, as well as
 4822 segmentation from co-stained images to improve segmentation. Notably, Baysor uses a

4823 Markov random field (MRF) to restrict the relationship between the spatially proximal
 4824 molecules. With each cell modeled with a Gaussian distribution, the entire dataset can
 4825 be regarded as a mixture of cell-specific distributions, which could be separable by
 4826 Bayesian mixture models (BMMs). Similarly, Sparcle utilizes the Dirichlet process
 4827 mixture model for initial cell-type identification and iteratively assigns each transcript
 4828 to cells by borrowing information from neighboring pixels (Prabhakaran, 2022).
 4829 Another method, ClusterMap, also leverages expression from the neighborhood to
 4830 compute a neighborhood gene composition (He et al., 2021b), and then formulates the
 4831 cell segmentation as a point pattern analysis problem, solvable by the density peak
 4832 clustering (DPC) algorithm.

4833 Following cell segmentation, cell-level analyses, such as differential expression
 4834 analysis and cell-cell interaction could be performed as in scRNA-seq. What's more,
 4835 further exploration of subcellular structures within cells becomes possible. For instance,
 4836 based on cell segmentation, ClusterMap can further segment cells into subcellular
 4837 structures including the nucleus and cytoplasm, using K-means clustering (Petukhov et
 4838 al., 2022). Bento, a toolkit for subcellular analysis of ST data, further enables the
 4839 identification of 5-class subcellular localization of RNA transcripts (Mah et al., 2022),
 4840 including nuclear, cytoplasmic, nuclear edge, cell edge, and none of the above.
 4841

4842 *Segmentation-free*

4843 The cell segmentation methods discussed above facilitate single-cell analysis on
 4844 imaging-based ST data. Nevertheless, challenges arise from technical noise such as
 4845 uneven intensity signals, and biological variation, including various cell sizes and
 4846 shapes and different cell densities. These factors can pose difficulties in achieving
 4847 accurate cell segmentation, potentially resulting in deviation in downstream analyses.
 4848 Therefore, several segmentation-free methods have been developed to enable robust
 4849 analysis without performing explicit segmentation. Most of the methods aim to assign
 4850 each molecule pixel to specific cell types rather than individual cells.

4851 To enable cell-type assignment for pixels, the authors of Baysor also provide a
 4852 segmentation-free approach (Petukhov et al., 2022). It is based on the assumption that
 4853 the neighboring RNA molecules are likely to stem from the same cell, collectively
 4854 reflecting the transcriptomics profile of the corresponding cell type. They compute a
 4855 neighborhood composition vector (NCV) for each transcript, effectively enhancing one
 4856 pixel's signal by leveraging information from its neighbors. The NCVs are
 4857 subsequently treated as 'pseudo-cells' for the downstream clustering and annotation
 4858 analyses. SSAM provides a similar solution (Park et al., 2021), which estimates the
 4859 mRNA signal for each pixel by borrowing its neighborhood's information. Differently,
 4860 they apply a Kernel Density Estimation (KDE) with a Gaussian kernel, differing from
 4861 Baysor, which gives equal weights to considered nearest neighbors.
 4862

4863 *Deciphering spatial distribution of cell types by integrating scRNA-seq*

4864 No matter whether tissue samples are profiled by single-cell or spatial transcriptomics,
 4865 cell-type annotation is always of great necessity to decipher cell compositions. The
 4866 annotation strategy designed for scRNA-seq, involving unsupervised clustering and

4867 cell-type inference based on expressed marker genes, seems applicable to the analysis
 4868 of ST data. Unfortunately, the attempt does not usually work owing to the limitations of
 4869 current ST technologies. First of all, for imaging-based targeted ST techniques, the
 4870 restricted selection of genes and the presence of read-out noise can hinder the
 4871 identification of unknown cell types. Second, for the low-resolution barcoding-based
 4872 ST data, the measurement of the mixture of multiple cells or cell types in each spot may
 4873 be averaged, potentially obscuring cell heterogeneity. Finally, for the high-resolution
 4874 barcoding-based ST data, the low detection efficiency also challenges both the manifest
 4875 clustering and proper cell-type annotation. As a result, in most cases, integrating ST
 4876 data with matched scRNA-seq becomes necessary to understand the cell-type
 4877 distribution. Generally, the integration can be accomplished by two approaches:
 4878 mapping or deconvolution.

4879

4880 *Cell Mapping*

4881 Cell Mapping includes two aspects: mapping pre-defined cell types to spatial locations
 4882 and mapping cells from scRNA-seq data to the tissue. The former transfers cell type
 4883 labels from scRNA-seq to spatial transcriptomics, while the latter predicts the spatial
 4884 locations for cells from scRNA-seq, which is also taken as the spatial reconstruction of
 4885 scRNA-seq in some cases.

4886 For cell-type mapping, one could calculate the enrichment score using
 4887 cell-type-specific gene signatures derived from scRNA-seq. This method has proven
 4888 effective in the analysis of microarray-based ST data of pancreatic ductal
 4889 adenocarcinomas. As for the imaging-based ST methods with limited genes, cell
 4890 segmentation methods mentioned above, such as pciSeq (Qian et al., 2020), JSTA
 4891 (Littman et al., 2021), and Baysor (Petukhov et al., 2022) could also allow cell-type
 4892 assignment when scRNA-seq is available. Alternatively, since these imaging-based ST
 4893 techniques could provide single-cell level expression after cell segmentation, existing
 4894 methods designed for single-cell data integration can be directly applied to integrate
 4895 single-cell resolution spatial data and scRNA-seq (Korsunsky et al., 2019; Stuart et al.,
 4896 2019; Welch et al., 2019; Peng et al., 2021b). For example, Seurat projects cells from
 4897 ST and scRNA-seq to the shared latent space by canonical correlation analysis (CCA)
 4898 (Stuart et al., 2019). With the cell pairs identified by mutual nearest neighbors (MNN)
 4899 as anchors, the cell-type labels from scRNA-seq could be transferred to spatial cells.
 4900 Similar integration can also be achieved by LIGER and Harmony. By leveraging the
 4901 common latent space and neighborhood information, these single-cell integration
 4902 methods can also predict the spatial expression of genes missed by ST, and strengthen
 4903 the original weak signals of ST-profiled genes.

4904 Spatial reconstruction of scRNA-seq, which predicts the spatial locations of cells
 4905 from expression with a few spatial landmark genes, was initially proposed before the
 4906 ST technologies boom. Earlier methods, such as Seurat (v1.0), models ISH reference
 4907 data with dozens of genes as a binarized expression map, and then probabilistically
 4908 infers a single cell's location by relating bimodal mixture models derived from
 4909 scRNA-seq to the binarized expression reference (Satija et al., 2015). Achim et al.
 4910 (Achim et al., 2015) and DistMap (Karaiskos et al., 2017) also utilize binarized ISH

reference but adopt different methods to calculate cell-location correspondence. Achim et al. designed a scoring scheme to assess the correspondence between a cell and each spatial location based on the gene specificity ratio in the given cell. DistMap calculates Matthew correlation coefficient (MCC) score using binarized single-cell gene expression and spatial reference, and then softly assigns cells to spatial locations (Achim et al., 2015). A recently developed method, Tangram is capable of aligning scRNA-seq to spatial transcriptomics measured by various technologies besides ISH-based data (Biancalani et al., 2021). By maximizing the correlation of gene expression shared by scRNA-seq and ST, Tangram could achieve a probability mapping matrix, which denotes the probability of finding every single cell in each spatial location.

Instead of scoring the cell-location correspondence, recent methods convert the problem of spatial reconstruction of scRNA-seq to a supervised learning problem or an optimization problem. For example, DEEPsc formulates the problem of mapping cells to spatial locations as a supervised classification problem by training a neural network-based classifier with spatial reference treated as scRNA-seq (Maseda et al., 2021). The sufficiently trained DEEPsc network takes the feature vector from a cell as input and predicts the cell's spatial origin according to the likelihood from different spatial locations. Another method, glmSMA frames cell mapping as a convex optimization problem (Gu and Liu, 2021). First, it employs Laplacian matrices to represent the location-to-location physical distance and cell-to-cell expression distance. By minimizing the differences between each cell's and corresponding locations' expression, glmSMA could finally find a mapping from cells in scRNA-seq to spatial locations in ST. SpaOTsc formulates cell mapping as an optimal transport problem, which aims to minimize the transport cost from cells to locations (Cang and Nie, 2020). The transport cost in SpaOTsc is measured majorly based on the gene expression dissimilarity across scRNA-seq and spatial reference and combines two penalty terms to handle the unbalanced sample sizes of two datasets and to preserve the structure within each dataset, respectively. Similarly, novoSpaRc adopts the framework of optimal transport, the core of which is the hypothesis that physically proximal cells share similar expression profiles (Nitzan et al., 2019). novoSpaRc measures the transport cost by the combination of location-to-location physical distance and cell-to-cell expression distance, both computed as the shortest path in their respective kNN graphs. By minimizing the transport costs, novoSpaRc finally obtains a certain mapping by which cells are mapped to locations with the original cell-cell correspondence preserved as much as possible, accounting for the above hypothesis. Notably, novoSpaRc also allows de novo reconstruction of scRNA-seq when reference ST data is not available. Most of the reconstruction methods are based on the assumption that the physical proximity could be reflected in expression similarity. However, the assumption cannot represent all the spatial distribution patterns of cells, which makes the inferred cell locations questionable.

4952

4953 *Cell type deconvolution*

4954 Deconvolution, which aims to estimate the exact cell-type proportions for each spatial

4955 location (i.e., spot or bead) is usually used in the integration of scRNA-seq and
 4956 low-resolution barcoding-based ST data, such as 10x Visium. For the high-resolution
 4957 barcoding-based ST techniques, such as Stereo-seq, the original pixel-level expression
 4958 is aggregated in a bin-based manner, and then each bin is treated as a new spatial unit
 4959 for deconvolution analysis. Current ST deconvolution methods can be basically divided
 4960 into four categories: regression, factorization, probabilistic modeling, and graph-based.

4961 Regression is one of the most popular methods developed for bulk RNA-seq
 4962 deconvolution. Due to the limited number of cells covered in each spot, the direct
 4963 application of bulk RNA-seq deconvolution methods on ST data will lead to noise from
 4964 unrelated cell types. To overcome this problem, spatialDWLS, an ST deconvolution
 4965 method based on the damped weight least square (DWLS) regression, adopts two
 4966 measures (Dong and Yuan, 2021). Firstly, the cell-type enrichment analysis is
 4967 performed before the accurate estimation of cell-type proportions to identify possible
 4968 cell types for each spot. Secondly, after the first round of deconvolution on the enriched
 4969 cell types, cell types predicted to have low proportions are removed to perform another
 4970 round of deconvolution.

4971 Methods based on regression highly rely on the selection of marker genes for each
 4972 cell type. Instead of performing regression on a cell-type-specific expression profile,
 4973 some methods propose to perform regression on a latent topic profile, which can be
 4974 decomposed from single-cell expression data by matrix factorization. For example,
 4975 NMFrreg, which was initially developed for cell-type annotation of Slide-seq, combines
 4976 non-negative matrix factorization (NMF) and non-negative least square (NNLS)
 4977 (Rodriques et al., 2019). It derives a basis gene-by-factor profile from pre-labeled
 4978 scRNA-seq using NMF, and then computes the factor loadings for each bead using
 4979 NNLS regression. With each factor linked to a cell type, the factor loadings serve as cell
 4980 type proportions. SPOTlight adopts a similar strategy, but uses a seeded NMF, in which
 4981 the combination of cell-type-specific marker genes and highly variable genes (HVG) is
 4982 used, and the factor-by-cell profile is initialized with the cell-cell-type belongingness
 4983 derived from scRNA-seq (Elosua-Bayes et al., 2021).

4984 Deconvolution can also be achieved with factorization alone. For instance,
 4985 STRIDE employs latent Dirichlet allocation (LDA), a topic modeling method, to derive
 4986 cell-type-associated topic profiles from scRNA-seq (Sun et al., 2022). Then, the
 4987 cell-type compositions of each spot can be inferred using the pre-trained topic model.
 4988 Stdeconvolve is also based on LDA but provides a reference-free solution (Miller et al.,
 4989 2022). CARD builds upon NMF, but takes spatial correlation between spots into
 4990 consideration by a conditional autoregressive (CAR) model, which makes CARD a
 4991 ‘spatial’ deconvolution method (Ma and Zhou, 2022).

4992 In addition to the intuitive regression or factorization-based methods, probabilistic
 4993 modeling approaches have been developed, assuming that the gene expression in a cell
 4994 or a spot follows a specific probabilistic model. For example, RCTD models the gene
 4995 expression in each location by a Poisson distribution and fits each spot as a linear
 4996 combination of individual cell types (Cable et al., 2022). Notably, RCTD also takes
 4997 platform-specific effects into account. Cell2location follows a similar concept, but uses
 4998 negative binomial (NB) distribution to model gene expression instead (Kleshchevnikov

et al., 2022). Likewise, Stereoscope utilizes the NB model, but it works on the complete set of genes rather than a set of selected marker genes (Andersson et al., 2020). DestVI also uses the NB distribution to model each gene's expression in a cell or a spot, with parameters encoded and decoded by neural networks (Lopez et al., 2022). Most importantly, DestVI not only estimates cell-type proportions, but also recovers cell-type-specific expression in each spot, which captures the continuous expression variation within cells of the same type.

Apart from DestVI, there are several other methods based on neural networks. DSTG first generates pseudo-ST data by randomly mixing cells from scRNA-seq and then constructs a link graph across spots from pseudo-ST and real-ST (Song and Su, 2021). With the link graph, which captures the intrinsic topological similarity between spots, semi-supervised graph convolutional network (GCN) is used to estimate cell-type proportions within each spot in real-ST. CellDART also generates pseudo-ST data – a virtual mixture of cells – but adopts the idea of adversarial domain adaptation (Bae et al., 2022). CellDART integrates two neural-network-based classifiers, in which the source classifier is trained to predict cell-type compositions, and the domain classifier is trained to discriminate real spots and pseudo spots. By iteratively updating two classifiers during training, the well-trained CellDART model could accurately estimate the cell-type proportions of each spot from real ST data. GraphST, another neural-network-based method, adopts a different strategy (Long et al., 2023). GraphST utilizes a graph contrastive self-supervised framework to reconstruct the gene expression for ST data by integrating spatial location information and local context. Using an autoencoder, GraphST can learn the latent representation of scRNA-seq separately. Based on the learned features, a cell-to-spot mapping probability matrix is trained through a contrastive learning mechanism, which can be combined with cell-type annotation of scRNA-seq to provide estimates of cell-type compositions for spots.

5027 Spatial domain identification

Cell type annotation for ST data could depict the spatial distribution of cell types in the tissue. More than discrete distribution, we are also interested in how the cell types are spatially organized to form the tissue architecture and execute functions. Intuitively, physically proximal cells, no matter from the same or different cell types, could constitute a spatial structure, which is usually termed spatial domain. Identification of spatial domains will help us understand the communication between cells within the domain and their biological functions (Jiang et al., 2024). In a sense, a spatial domain can be regarded as a cluster of cells with specific spatial patterns. The standard Louvain clustering method for scRNA-seq is based on the graph built upon gene expression similarity, which does not consider spatial information and is not directly applicable here. Some spatial clustering methods modify graph-based clustering algorithms to incorporate the spatial information. For example, stLearn utilizes Louvain or K-means for global clustering and performs local clustering to find spatially separated sub-clusters or merge spatially proximal singleton spots by considering physical distances (Pham et al., 2020). Another method, MULTILAYER applies Louvain

clustering on the gene-pattern co-expression graph (Moehlin et al., 2021). At first, MULTILAYER detects expression patterns for overexpressed genes by an iterative agglomerative strategy. A gene expression pattern here is defined as a region with the gene overexpressed in multiple contiguous locations. Then MULTILAYER constructs a graph where nodes represent previously detected gene patterns, and edges represent the similarity between gene patterns (i.e., gene co-expression degree). Finally, the Louvain algorithm is implemented to partition the gene co-expression patterns into multiple tissue communities.

Instead of incorporating spatial information in indirect ways, many spatial clustering methods encode the information of spatial proximity in an MRF, in which the spatial dependence is formulated by the Potts model. Zhu et al. developed smfishHmrf, which applied hidden Markov random fields (HMRF) to the identification of spatial domains from seqFISH data (Zhu et al., 2018b). They first construct a neighborhood graph to represent the spatial relationship between cells, in which the Markov property keeps only relationships between immediately neighboring nodes. Then they model each cell's domain state by a joint probability distribution, which considers both the cell's gene expression and the domain states of neighboring cells. By solving parameters for the equilibrium of the field using expectation-maximization (EM), smfishHmrf enables the detection of spatial domains with spatially coherent gene expression. BayesSpace adopts a fully Bayesian statistical model with an MRF to ensure spots from the same cluster are closer to each other physically (Zhao et al., 2021a). By using Markov chain Monte Carlo (MCMC) and a fixed precision matrix across different clusters, BayesSpace is able to stably estimate model parameters, identify spatial clusters, and even enhance the resolution of spatial transcriptomics. Given that MCMC is computationally intensive and the fixed smoothness parameters may limit the performance in different ST datasets, Yang et al. proposed SC-MEB to enable both efficient computation and adjustable smoothness parameters (Yang et al., 2022). In particular, they applied an efficient iterative-conditional-mode-based expectation-maximization (ICM-EM) scheme to estimate parameters, and selected the cluster number by the modified Bayesian information criterion (MBIC). The above MRF-based methods all assume the hidden cell states to be discrete, limiting our understanding of spatial dependency among cells. In contrast, SPICEMIX integrates NMF into HMRF, in which observed gene expression is modeled as linear mixtures of latent factors, and the mixing weights of latent factors are regarded as the hidden cell states (Chidester et al., 2021). SPICEMIX, to understand it in another way, provides a method of dimension reduction for ST data by considering spatial information, which could be the foundation of downstream clustering. Based on the inferred cell states, hierarchical clustering is further applied by SPICEMIX to define categorical cell types.

Cell type clustering and spatial domain identification can be treated as two separate tasks in ST data analysis. Most of the methods we discussed above focus on identifying spatial domains, except SPICEMIX, where spatial clustering is intended to infer cell types without integration with scRNA-seq. Another method, FICT aims to infer cell types in FISH-based spatial transcriptomics by spatial clustering (Teng et al., 2022). Specifically, FICT models the expression of a cell by a cell-type-specific

5087 Gaussian distribution and models the relationship between the cell and its neighboring
 5088 cells by a multinomial distribution. FICT is capable of assigning cell clusters by
 5089 maximizing the joint probabilistic likelihood. Similarly, BASS also models the gene
 5090 expression in a cell by a cell-type-specific normal distribution, but meanwhile, it
 5091 models the cell type belongingness by a domain-specific categorical distribution (Li
 5092 and Zhou, 2022). With such a hierarchical probabilistic framework, BASS enables
 5093 simultaneous cell type clustering and spatial domain detection.

5094 Spatial transcriptomics can be naturally regarded as a spot-spot graph, which is
 5095 suitable to be fed to graph-based neural networks. Many graph-neural-network-based
 5096 (GNN-based) methods have been developed to learn the low-dimensional latent
 5097 representations from spatial transcriptomics by integrating gene expression and spatial
 5098 information (Cang et al., 2021; Hu et al., 2021a), which can facilitate the downstream
 5099 analyses such as spatial domain identification and detection of spatially variable genes.
 5100 For example, SpaGCN applies a graph convolutional network (GCN) to integrate
 5101 multiple sources of information, including gene expression, spatial locations, and
 5102 histology (Hu et al., 2021a). Firstly, a graph is built to represent the relationship among
 5103 spots, where nodes represent spots, and the distances of the edges are calculated by
 5104 converting histology image features to the third ‘z’ coordinate and combining it with
 5105 the spots’ original spatial coordinates (x, y). Then a convolutional layer is applied to
 5106 aggregate gene expression from neighboring spots in the graph. Based on the
 5107 aggregated gene expression, an unsupervised iterative clustering algorithm is then
 5108 implemented to identify clusters (i.e., spatial domains).

5109 Other methods introduce additional mechanisms into the basal GCN. As we
 5110 discuss in the section on cell-type deconvolution, GraphST applies a graph contrastive
 5111 self-supervised framework to learn the spatial latent representations for ST data by
 5112 combining gene expression with spatial location information and local context
 5113 information (Long et al., 2023). Another method, SpaceFlow integrates a deep graph
 5114 infomax (DGI) framework into the GCN encoder (Ren et al., 2022). In addition to a
 5115 spatial expression graph (SEG) built from spatial transcriptomics, SpaceFlow also
 5116 constructs an expression permuted graph (EPG) by randomly permuting expression.
 5117 The two graphs are both fed to a graph convolutional encoder to get the
 5118 low-dimensional embeddings, and the DGI enables the encoder to distinguish
 5119 embeddings of SEG from those of EPG through a discriminator loss. Some methods
 5120 take autoencoders for spatial embedding. For instance, SEDR employs a deep
 5121 autoencoder network to learn a low-dimensional latent representation for gene
 5122 expression, which is later integrated with spatial information using a variational graph
 5123 autoencoder (VGAE) (Fu et al., 2021a). STAGATE introduces an attention mechanism
 5124 to the autoencoders, enabling adaptive learning of the edge weights (i.e., spot similarity)
 5125 (Dong and Zhang, 2022). stMVC constructs a more comprehensive learning framework
 5126 (Zuo et al., 2022). To be specific, stMVC first learns visual features from histology
 5127 images through data augmentations and contrastive learning. Then semi-supervised
 5128 graph attention autoencoders (SGATE) are used to learn view-specific representations
 5129 based on the extracted visual features and spatial gene expression independently and
 5130 integrate two graphs via an attention mechanism. The attention-based multi-view graph

5131 collaborative learning model proposed by stMVC finally learns a more robust
 5132 representation of ST data.

5133 Due to ST data's essence of spatial signals, some methods translate the problem of
 5134 spatial domain identification to the classic image segmentation problem. RESEPT uses
 5135 GNN to learn a three-dimensional embedding from a spot-spot graph, with gene
 5136 expression treated as the nodes' attributes and physical adjacency revealed by edge
 5137 connectivity (Chang et al., 2022). The three-dimensional embedding of each spot is
 5138 transformed to an RGB scale so that a previous CNN designed for semantic
 5139 segmentation can be directly applied to segment spatial domains. Another method,
 5140 Vesalius adopts a similar RGB embedding strategy, but through dimension reduction by
 5141 UMAP rather than neural networks (Martin et al., 2022).

5142

5143 *Detection of spatially variable genes and gene expression patterns*

5144 Highly variable genes (HVG) play a critical role in dimension reduction and
 5145 subsequent cell clustering in the analysis of scRNA-seq. In spatial transcriptomics, the
 5146 identification of spatially variable genes (SVG) is also important to characterize the
 5147 functional organization in complex tissues. To identify SVG is to find genes showing
 5148 great variation in space. HVG detection in scRNA-seq only considering the high
 5149 variance but ignoring spatial information, cannot be directly applicable in the SVG
 5150 identification. Various computational methods have been proposed to detect SVG from
 5151 spatial transcriptomics. Some of the methods identify SVG based on segmented spatial
 5152 domains. For example, SpaGCN first identifies spatial domains by integrating multiple
 5153 sources of information as we discussed above and then defines the neighboring domain
 5154 for each identified domain (Hu et al., 2021a). The spatially variable genes are
 5155 determined by identifying differentially expressed genes between each target domain
 5156 and the corresponding neighbor domain using the Wilcoxon rank-sum test. Instead of
 5157 relying on spatial domain identification, most methods directly incorporate spatial
 5158 information into the models to study the spatial variance of gene expression. According
 5159 to the core models, methods could be generally divided into three categories: methods
 5160 based on statistical modeling, graphs, and other principles.

5161

5162 *Based on statistical modeling*

5163 Trendsseek models the spatial expression as marked point processes, where the spatial
 5164 locations are considered as a two-dimensional point process, and the locations' expressions
 5165 are treated as marks (Edsgard et al., 2018). For a given gene and a specified
 5166 distance, the dependency between the spatial distribution of points and their marks is
 5167 evaluated for all point pairs at the distance. The dependency assessment could be
 5168 achieved by four summary statistics. Stoyan's mark-correlation, mean-mark,
 5169 variance-mark, and mark-variogram. The summary statistics will remain constant when
 5170 the marks and the distribution of marks are independent, but if they are dependent, the
 5171 statistics will vary across different distances. Significance is estimated by permuting
 5172 the expression values, and the smallest p-value among different distances is regarded as
 5173 the significance of the gene. scGCO also utilizes the marked point process to model
 5174 spatial gene expression but integrates HMRF into the model (Zhang et al., 2022b). For

5175 each gene, scGCO segments the graph representation by a graph cuts algorithm. The
 5176 segments can be used as the candidate regions to test the expression's dependence on
 5177 the spatial locations under the complete spatial randomness framework, where the
 5178 distribution of points in 2D space is modeled as a homogeneous Poisson process.

5179 In addition to the marked point process, many methods utilize the Gaussian
 5180 process (GP) to model spatial gene expression. GP is a collection of random variables
 5181 indexed by time or space, in which any finite collection of these random variables has a
 5182 multivariate normal distribution. GP is widely used in geostatistics and has been
 5183 applied in modeling spatial transcriptomics. For example, SpatialDE, based on
 5184 Gaussian process regression, models each gene's variability with two components:
 5185 spatial and non-spatial variance terms. The ratio of these terms can be calculated to
 5186 quantify the spatial variability (Svensson et al., 2018). Statistical significance could be
 5187 estimated with a log-likelihood test by comparing the likelihood of the full model with
 5188 the null model without spatial covariance. SpatialDE could further identify genes with
 5189 different types of spatial variation, including linear or periodic patterns, by comparing
 5190 the full model fitted with a linear or periodic (i.e., cosine) covariance function with that
 5191 of the Gaussian kernel. To meet the assumption of Gaussian distribution, SpatialDE
 5192 employs a two-step normalization. Specifically, SpatialDE uses a variance-stabilizing
 5193 transformation method, known as Anscombe's transformation, to transform the
 5194 NB-distributed raw counts followed by regression of log total counts. Gpcounts also
 5195 builds on Gaussian process regression, but adapts it by fitting the spatial counts by NB
 5196 or zero-inflated negative binomial (ZINB) distribution rather than Gaussian
 5197 distribution (BinTayyash et al., 2021). Similarly, BOOST-GP models gene read counts
 5198 through a ZINB distribution but adopts a Bayesian framework to infer the parameters
 5199 (Li et al., 2021b). Another method, SPARK, employs the generalized linear spatial
 5200 model (GLSM) with GP modeling the spatial relationships between spatial locations
 5201 and Poisson distribution modeling the expression count data (Sun et al., 2020a).
 5202 Moreover, SPARK provides a more powerful statistical method to control type I errors,
 5203 which computes p-values for each parameterized kernel separately and combines them
 5204 with the Cauchy combination rule.

5205 With the development of ST techniques, previous methods need to be modified to
 5206 adapt to large-scale spatial transcriptomics data of high sparsity. Based on SPARK,
 5207 SpatialDE2 improves the computational efficiency by replacing the Cauchy
 5208 combination with the omnibus test and introducing GPU acceleration of Tensorflow
 5209 (Kats et al., 2021). In order to reduce the computational complexity and physical RAM
 5210 requirement, the authors of SPARK proposed a scalable non-parametric test method,
 5211 SPARK-X (Zhu et al., 2021b). To be specific, SPARK-X builds on a non-parametric
 5212 covariance test framework, in which two covariance matrices are calculated to measure
 5213 the expression similarity and spatial proximity, respectively. Then identifying genes
 5214 with specific spatial trends is converted to testing the dependence between gene
 5215 expression and spatial locations. Another method, SOMDE incorporates the
 5216 self-organizing map (SOM) neural network into the Gaussian process regression
 5217 framework of SpatialDE (Hao et al., 2021a). SOMDE condenses the original spatial
 5218 locations into SOM nodes with the spatial expression pattern and the topological

structure preserved. The original spatial expression is then aggregated to form the node-level gene meta-expression, which significantly reduces the size of the covariance matrix, and thus increases the computational efficiency.

5222

5223 *Based on graph representation*

5224 As discussed in the section on spatial domain identification, the spatial expression can
 5225 be represented by a graph. Some graph-based methods have been demonstrated to be
 5226 successful in SVG identification. The graph Laplacian score, commonly used for
 5227 graph-based feature selection, can be applied to identify spatially variable genes from
 5228 graphs. GLISS, for instance, first builds a mutual nearest neighbor graph and computes
 5229 a Laplacian score for each gene to measure its locality-preserving power (i.e., its
 5230 association with local structures) (Zhu and Sabatti, 2020). A low Laplacian score,
 5231 within a fixed graph, indicates that similarity of gene expression occurs in close
 5232 locations, whereas large variation occurs in more distant locations (He et al., 2005).
 5233 The statistical significance of each gene is estimated by permuting expression with the
 5234 graph fixed. RayleighSelection proposed combinatorial Laplacian scores with the
 5235 graph-based representation extended to the simplicial complex representation of spatial
 5236 expression data (Govek et al., 2019). Apart from vertices and edges included in graphs,
 5237 simplicial complexes also contain higher-dimensional elements such as triangles and
 5238 tetrahedrons, which could capture more complex relations of the data. Accordingly, the
 5239 combinatorial Laplacian score facilitates the identification of genes with more complex
 5240 spatial structures.

5241 Some methods introduce spatial gridding into the ordinary graph representation to
 5242 simplify or optimize the spatial structure. singleCellHaystack, a spatial-gridding-based
 5243 approach, was initially developed to predict differentially expressed genes from
 5244 low-dimensional spaces learned from scRNA-seq, independent of cell clustering
 5245 (Vandenbon and Diez, 2020). It can also be applied to the SVG identification of spatial
 5246 transcriptomics data using the natural 2D or three-dimensional space.
 5247 singleCellHaystack first divides the multi-dimensional space into grids and defines grid
 5248 points, which are used to estimate the reference distribution of cells in the space. Then
 5249 for each gene, singleCellHaystack classifies all cells into detected and undetected
 5250 groups according to the binarized expression and estimates the cell distribution
 5251 separately. Kullback-Leibler Divergence is subsequently calculated to measure the
 5252 gene's divergence by comparing it with the reference cell distribution, and the
 5253 significance is evaluated by permutation test. MERINGUE is another method based on
 5254 spatial gridding (Miller et al., 2021). It starts by constructing the neighborhood
 5255 adjacency relationships using Voronoi tessellation, which is also used for the
 5256 construction of graph representation in scGCO. Compared to the k-nearest neighbor or
 5257 k-mutual-nearest neighbor, Voronoi tessellation adapts to varying neighborhood sizes
 5258 and distances, offering better stability in tissues with diverse cell types and
 5259 non-uniform densities. Then MERINGUE computes Moran's I for each gene to
 5260 measure the spatial auto-correlation, which indicates the expression correlation among
 5261 spatially adjacent locations. Giotto also provides a spatial gridding-based method,
 5262 BinSpect (Dries et al., 2021b). Similarly, BinSpect relies on Voronoi tessellation to

5263 determine the neighborhood relationships. Instead of Moran's I, BinSpect adopts the
 5264 statistical enrichment analysis. For each gene, BinSpect binarizes the expression using
 5265 k-means clustering with $k = 2$ or simple thresholding on rank. Next, a contingency table
 5266 is calculated to reflect the expression dependency between neighboring locations. A
 5267 Fisher exact test is then employed to obtain an odds ratio and the corresponding p-value.
 5268 If a gene is found to be significant, it tends to be highly expressed in the neighboring
 5269 locations.

5270

5271 *Based on other principles*

5272 In addition to methods rooted in statistical models or graph representation, there are
 5273 approaches using entirely different principles. Sepal proposed a unique strategy
 5274 founded on the diffusion theory, which regards the observed gene expression profile as
 5275 the outcome of transcript diffusion. Within the framework of simulation, sepal assumes
 5276 that it will take more time for transcripts to form a structured pattern than to reach a
 5277 homogeneous random state. Hence, inferring the structured degree of gene expression
 5278 patterns is converted to measuring the diffusion time in the simulation system. Another
 5279 method, SPADE focuses on identifying important genes associated with morphological
 5280 features (Bae et al., 2021). SPADE extracts latent image features from histological
 5281 images by utilizing a convolutional neural network (CNN). Principal component
 5282 analysis (PCA) is then performed on the high-dimensional features to summarize the
 5283 spatial distribution patterns of image features. SPADE uses a linear model to discover
 5284 genes correlated with the image patterns (i.e., PCs), which have been demonstrated to
 5285 exhibit specific spatial trends.

5286 To model the spatial variation of gene expression, the methods discussed above
 5287 only consider the relative distance between locations, ignoring the variation along
 5288 specific directions. SPATA offers an option for users to manually define a trajectory
 5289 axis according to prior knowledge (Kueckelhaus et al., 2020). For each gene, multiple
 5290 functions are fitted to model the spatial variation patterns along the predefined spatial
 5291 axis, including linear, logarithmic, or gradient ascending/descending, one-, or
 5292 multiple-peak functions. Among all the functions, the best-fitting one is selected to
 5293 represent the gene's dynamics by comparing the summed residuals.

5294 After spatially variable genes are identified, some methods further determine
 5295 archetypal gene patterns through clustering. By an extended Gaussian mixture model
 5296 with a spatial prior on cluster centroids, SpatialDE conducts clustering to group SVGs
 5297 with similar spatial expression patterns (Svensson et al., 2018). Similarly, SPARK
 5298 implements a hierarchical clustering algorithm to classify detected variable genes into
 5299 different categories (Sun et al., 2020a). Instead of constructing similarity matrices
 5300 based on expression, MERINGUE derives a cross-correlation matrix by computing a
 5301 spatial cross-correlation index, which is a modification of Moran's I auto-correlation
 5302 for each pair of genes. This forms the basis of hierarchical clustering. GLISS fits a
 5303 spline model on the latent structure, where each gene can be represented by the fitted
 5304 spline coefficients and genes with similar gene patterns will share similar coefficients.
 5305 Compared to expression-based similarity, computing gene-gene similarity based on
 5306 spline coefficients could reduce correlation unrelated to spatial variation. Then GLISS

5307 performs spectral clustering on the coefficients to cluster genes into groups.

5308

5309 *Pseudo-time trajectory analysis*

5310 From scRNA-seq or ST data, we capture only a snapshot of the cellular gene
 5311 expression. The above spatial domain detection or SVG identification enables us to
 5312 study the transcriptional dynamics by space in a discrete or continuous way,
 5313 respectively. Previous efforts in pseudotime analysis of scRNA-seq have provided us
 5314 with opportunities to reconstruct cell state trajectories from expression data alone. The
 5315 additional spatial information brought by ST expands the original pseudotime analysis
 5316 by introducing the dimension of space.

5317 Direct application of single-cell pseudotime methods on ST data may cause cell
 5318 trajectory to be continuous with time but discontinuous in space. To address the
 5319 problem, stLearn adapts the original pseudotime algorithm by incorporating spatial
 5320 information (Pham et al., 2020). stLearn first utilizes the diffusion pseudotime (DPT)
 5321 algorithm to predict pseudotime from gene expression. Then it computes a
 5322 pseudo-space-time distance (PSTD) matrix by combining differences in
 5323 expression-based pseudotime and spatial distance, with a weight to balance between
 5324 them. Based on the PSTD matrix, stLearn constructs a directed graph and applies a
 5325 minimum spanning tree algorithm to determine branches (i.e., to infer cell trajectories).

5326 Instead of relying on the initial pseudotime trajectories inferred only from gene
 5327 expression, several methods emerged to predict the cell trajectories from combined
 5328 expression and spatial information. SpaceFlow, which has been discussed in the section
 5329 on spatial domain identification, provides a deep learning framework to learn
 5330 low-dimensional embeddings from ST data (Ren et al., 2022). The embeddings
 5331 produced by SpaceFlow could be used to calculate the pseudo-Spatiotemporal Map
 5332 (pSM) using the DPT algorithm, facilitating the integrative reconstruction of
 5333 spatiotemporal trajectories from ST data. Consequently, the spatiotemporal order
 5334 generated by SpaceFlow maintains consistency in both space and pseudotime.

5335

5336 *Cell-cell communication and gene-gene interaction*

5337 Through the aforementioned analyses, we could get a basic understanding of the spatial
 5338 distribution of cell types and the expression variations in space. However, the
 5339 organization of cells or cell types, as well as the regulation of genes to generate such
 5340 spatial patterns, remain elusive. Many studies have reported that cellular behavior can
 5341 be shaped by cell signaling pathways from the environment. Spatial transcriptomics
 5342 offers a unique opportunity to investigate cell-cell communications within the
 5343 preserved microenvironment. Several methods have been proposed to explore spatial
 5344 dependence between cells from ST data, among which the most intuitive is to study the
 5345 proximity or the co-localization of different cell types. Giotto, for instance, adopts a
 5346 random permutation strategy to identify the enriched cell-type pairs (Dries et al.,
 5347 2021b). With the structure of the neighborhood network fixed, cell-type labels are
 5348 shuffled among the nodes to form random neighboring relationships. In this way, the
 5349 ratio of observed-over-expected frequencies between two cell types is determined, and
 5350 the corresponding enrichment significance can be estimated. spicyR, originally devised

for spatial analysis of *in situ* cytometry, defines a score to measure the degree of cell-type co-localization (Canete et al., 2022). With the spatial distribution of cells modeled by the marked point process, spicyR applies a K-function or variance stabilized K-function (i.e., L-function) to quantify the co-localization between two cell types within a specific distance. Recently, Cang et al. developed COMMOT, based on a collective optimal transport method, to handle complex molecular interactions and spatial constraints for inferring paracrine-dependent cell-cell communication in spatially resolved transcriptomics. (Cang et al., 2023)

Beyond observed co-localization of cell types, the spatial dependence among cells can be more complicated, which needs to be modeled by more complex methods. NCEM reconciles variance attribution and cell-cell communications in a node-centric expression model (Fischer et al., 2021). NCEM first uses the graph structure to enforce a neighborhood constraint on the cell communications. With the provided cell-type labels, NCEM applies a function to fit a cell's observed gene expression by its cell type and its spatial context. To accommodate the complexity of the spatial dependencies in different scenarios, NCEM provides three models, including the linear, nonlinear, and generative latent variable models, which are implemented by linear regression, nonlinear encoder-decoder GNN, and conditional variational autoencoder, respectively. By modeling the dependencies of the molecular states of the target cell (i.e., receiver) on the neighborhood (i.e., sender), NCEM can also determine the directionality of the sender-receiver signaling.

Instead of modeling the entire expression profile's dependence on cell-cell communications, several methods quantify the effect of cell-cell interactions on expression for each gene individually. For instance, SVCA models the expression of a target gene across cells with the Gaussian process model and decomposes the gene's variability into three components, including intrinsic effects, environmental effects from unmeasured spatial variables, and cell-cell interaction effects from neighboring cells (Arnol et al., 2019). In this manner, the fraction of variance explained by each term can be estimated for each gene, and the biologically relevant genes participating in cell-cell interactions can be identified. MISTy designs a multi-view framework to account for the expression of individual genes, where cell-cell interactions from different spatial contexts are modeled in different views (Tanevski et al., 2022). Similar to SVCA, MISTy includes intraview, juxtaview, and paraview, which correspond to intrinsic effects from gene expression of other genes in the same location, effects from immediate neighbors, and effects from the tissue architecture (i.e., cells within a radius of the specified cell), respectively. By analyzing each predictor gene's importance to the target gene in each view, the effects from different spatial contexts can be explainable for the gene pair of interest.

SVCA and MISTy can model the gene-gene relations, and discover genes associated with cell-cell interactions, but neither of them can identify explicit gene-gene interaction pairs. Yuan and Bar-Joseph developed GCNG, a GCN-based supervised computational framework, to predict gene-gene interactions (Yuan and Bar-Joseph, 2020). GCNG takes the graph representation of spatial neighborhood as input, as well as the normalized expression of candidate gene pairs. The output will be

the classification of the interacting or non-interacting gene pairs. To enable supervised learning, known ligand-receptor interactions from a curated list are labeled as positive pairs, and randomly selected ligand-receptor pairs are labeled as negative data. With a five-layer GCN structure, GCNG could predict new gene-gene interactions in the studied ST dataset. However, GCNG cannot inform the cell types where interactions occur, and cannot focus on interaction inference within specific local regions of interest either. To address these limitations, some methods rely on the co-expression of ligands and receptors by taking cell-type locations into consideration (Pham et al., 2020; Dries et al., 2021b; Garcia-Alonso et al., 2021). For example, MERINGUE further constrains the spatial cross-correlation calculation between pairs of genes to the curated ligand-receptor pairs and two cell types of interest (Miller et al., 2021). Garcia-Alonso et al. upgraded their CellphoneDB to v3.0, which identifies ligand-receptor pairs within specific microenvironments where cell types of interest are co-localized (Garcia-Alonso et al., 2021). Similarly, based on the cell-type proximity analysis in the previous step, Giotto defines a ligand-receptor interaction score by calculating the weighted average expression of ligands and receptors in the cell subsets of interacting cell types.

5412

5413 *Integrative analysis of spatial data*

5414 With increased throughput and decreased costs, some studies generate ST slides from
 5415 multiple individuals to perform large-scale analysis. Some other studies produce a
 5416 series of ST slides from multiple adjacent layers of the tissue, enabling a global view of
 5417 the whole tissue. Conducting separate analyses on individual ST slides may diminish
 5418 the power of multiple samples. Thus, integration methods are required to perform a
 5419 joint analysis of multiple samples. Moreover, with additional information such as
 5420 morphologies provided, spatial transcriptomics should be integrated with other
 5421 modalities to comprehensively characterize the tissue. In this section, we will review
 5422 computational methods for the integration of multiple samples and the integration of
 5423 spatial data from various modalities.

5424

5425 *Multi-sample integration*

5426 The core of multi-sample integration involves placing multiple samples in the same
 5427 space, referred to as common coordinate framework (CCF). The coordinate system
 5428 encompasses two facets. On one hand, CCF can represent the natural
 5429 three-dimensional space, in which multiple plane slides are aligned and stacked to
 5430 provide a stereoscopic view of tissues. On the other hand, high-dimensional spatial
 5431 measurements of location from multiple samples could be projected into a shared
 5432 low-dimensional space for integrative analyses such as joint spatial domain
 5433 identification.

5434 Some methods have been developed to align multiple sequential slides from the
 5435 same tissue. PASTE formulates the multi-slide alignment as an optimal transport
 5436 problem, which computes the probabilistic alignment based on both gene expression
 5437 and spatial information (Zeira et al., 2022). By minimizing the transport cost, PASTE
 5438 could achieve a mapping that maximizes gene expression similarity between aligned

5439 locations across slides while preserving spatial structure within a slide. PASTE can
 5440 align multiple sequential slides from the same tissue, but cannot be applied to the
 5441 integration of slides from different time points. Andersson et al. proposed a method,
 5442 eggplant, which is a landmark-based method to project multiple slides into the common
 5443 reference (Andersson et al., 2021a). First, eggplant projects the measured spatial
 5444 locations to the reference by making the distance between landmarks conserved before
 5445 and after transformation. Next, eggplant applies the Gaussian Process Regression to
 5446 learn the relationship between the gene expression and the distance to the landmarks for
 5447 all landmark-excluded locations, allowing prediction of gene expression for each
 5448 location in the reference. With the strategy of location transformation combined with
 5449 expression prediction, multiple slides at different time points or from different
 5450 individuals could be transferred to the same reference for integrative analysis. However,
 5451 eggplant necessitates not only the selection of landmark locations but also the
 5452 definition of reference, which is usually a canonical structure representing the tissue
 5453 domain. Both requirements limit eggplant's application on more complicated tissues
 5454 such as tumors. To address this issue, Jones et al. developed GPSA, which is also based
 5455 on the Gaussian process model (Jones et al., 2022a). GPSA constructs a two-layer
 5456 Gaussian process framework, where the first layer maps the measured spatial locations
 5457 to a common coordinate system, and the second layer describes the spatial gene
 5458 expression within this system. Compared to eggplant, GPSA could iteratively estimate
 5459 the common coordinate system *de novo*, but it also offers an option for template-based
 5460 alignment with a pre-defined common coordinate system by fixing one slide.

5461 Instead of mapping spatial locations from multiple slides to the CCF in the natural
 5462 3D space, several methods focus on projecting multiple samples to a shared
 5463 low-dimensional space. In this case, integration methods should be capable of
 5464 removing unwanted variations from different batches and preserving the meaningful
 5465 biological variations as in scRNA-seq. But different from single-cell integration
 5466 methods, ST integration methods should take into account spatial information. Liu et al.
 5467 proposed PRECAST, a unified and principled probabilistic model, to jointly estimate
 5468 low-dimensional embeddings and perform spatial clustering across multiple tissue
 5469 slides (Liu et al., 2022). PRECAST performs dimension reduction on the normalized
 5470 gene expression using the intrinsic conditional autoregressive (CAR) model, which
 5471 could preserve the original spatial dependence among neighbors in the
 5472 low-dimensional space. The resulting latent low-dimensional embedding could be
 5473 further employed to perform spatial clustering using an HMRF model. As we
 5474 mentioned above, BASS enables multi-scale analysis for simultaneous cell type
 5475 clustering and spatial domain detection. It also allows the multi-sample integration
 5476 analysis by jointly modeling the Harmony-corrected spatial transcriptomics with a
 5477 hierarchical Bayesian framework. Another method, MAPLE proposed a hybrid
 5478 framework for joint spatial clustering of multiple sections, following the spatially
 5479 aware low-dimensional embedding learning via a GCN-based model (Allen et al.,
 5480 2022).

5481

5482 *Multi-modal integration*

5483 As we discussed above, single-cell and spatial transcriptomics are usually integrated to
 5484 decipher the spatial distribution of cell types through cell mapping or cell-type
 5485 deconvolution. Among the integration methods we reviewed, Tangram stands out by
 5486 enabling the mapping of data from other modalities onto the spatial transcriptomics
 5487 through integration with multi-modal single-cell data. For example, once the single
 5488 cells from SHARE-seq are mapped to spatial locations by gene expression similarity,
 5489 the spatial patterns of chromatin accessibility can be unveiled.

5490 Considering that many ST technologies provide corresponding histological
 5491 images, many computational methods leverage the additional image information to
 5492 enhance the analytic performance at each step. For example, stLearn leverages
 5493 morphological similarity to normalize the expression data, thereby reducing the impact
 5494 of technical noise of dropouts (Pham et al., 2020). spaGCN takes the histology image
 5495 features into consideration when calculating spot-spot distances to construct a graph for
 5496 spatial transcriptomics (Hu et al., 2021a). stMVC employs graph networks with the
 5497 attention mechanism to integrate multi-source information including histological
 5498 features, and ultimately learns the low-dimensional embedding of ST data (Zuo et al.,
 5499 2022). Likewise, methods such as conST (Zong et al., 2022) and MUSE (Bao et al.,
 5500 2022) also use deep learning architectures to integrate cell morphologies and
 5501 transcriptional states for joint representation. Instead of the complex deep
 5502 learning-based mechanism, SPADE directly associates the spatial variance of gene
 5503 expression with the spatial distribution patterns of image features using a linear
 5504 regression model (Bae et al., 2021).

5505 In addition to facilitating the analysis of spatial transcriptomics, the histological
 5506 images could also be used to predict spatial gene expression. Many methods have been
 5507 developed to address such a problem. To overcome the limitation of low resolution in
 5508 some barcoding-based ST technologies, Bergenstråle et al. proposed a deep generative
 5509 model to infer the super-resolved expression maps from high-resolution histology
 5510 images, both within and between the originally measured locations (Bergenstrahle et al.,
 5511 2022). Rather than focusing on improving the resolution of spatial gene expression,
 5512 some methods generalize the spatial transcriptome prediction to the histopathology
 5513 images without matched expression data. For example, He et al. introduced a deep
 5514 learning algorithm, ST-Net, to capture gene expression heterogeneity by combining
 5515 spatial transcriptomics and histology images (He et al., 2020a). With the model trained
 5516 with a BRCA spatial transcriptomics dataset comprising 68 ST slides of breast tissue
 5517 sections, it can predict the spatially resolved transcriptome of other breast cancer
 5518 datasets directly from histology images. However, ST-Net does not account for spatial
 5519 dependencies between spots. HisToGene employs a modified Vision Transformer
 5520 model to enable the prediction of spatial gene expression with the spot dependency
 5521 considered (Pang et al., 2021). Building upon HisToGene, Hist2ST additionally
 5522 includes a Conv mixer module to capture the internal relations of 2D vision features
 5523 within image patches (Zeng et al., 2022b).

5524

Applications

5525 The recent and rapid progress in spatial transcriptomics has promoted its widespread

application across various biological systems. ST techniques have been instrumental in spatially characterizing the cell states of healthy tissues, and some of them aim to decipher the spatial architecture of tissues at specific developmental stages. Notably, among the tissues, the nervous system has been a focal point of investigation. Numerous studies have made substantial contributions to constructing detailed spatial atlases for the brain. Moreover, ST techniques have proven invaluable in exploring the microenvironments of injured or diseased tissue, including mouse lungs infected with virus (Boyd et al., 2020), human hearts with myocardial infarction (Kuppe et al., 2022), as well as a range of different tumor types (Ji et al., 2020a; Wu et al., 2021a; Wu et al., 2021b; Wu et al., 2021d; Qi et al., 2022). Here we review the applications of ST in three main fields, encompassing the development and homeostasis of healthy tissues, neuroscience, and the tumor microenvironment.

5539

5540 *Development and homeostasis of healthy tissue*

5541 Most of the studies utilize mouse models to investigate the development of early
5542 mammalian embryos. Spatial atlases have been established for several stages of mouse
5543 embryonic development. Peng et al. focused on lineage differentiation and
5544 morphogenesis at the post-implantation stages (Peng et al., 2019). Geo-seq was applied
5545 to profile cell populations at pre-selected positions in all germ layers from
5546 pre-gastrulation (embryonic day 15.5) to late gastrulation (E7.5). The study unveiled
5547 the dynamic molecular regulation of lineage specification and tissue patterning in time
5548 and space. Moreover, they also uncovered the pivotal role of Hippo/Yap signaling
5549 during germ-layer development. To further explore the cell fate decisions in the early
5550 organogenesis at the end of gastrulation, Lohoff et al. performed seqFISH on multiple
5551 sagittal sections collected from mouse embryos at E8.5-8.75 (Lohoff et al., 2022). Due
5552 to the limited number of target genes, they integrated seqFISH with existing single-cell
5553 transcriptome atlases to enable genome-wide imputation. By utilizing the generated
5554 spatial single-cell map, the authors revealed spatial patterns of gene expression
5555 corresponding to dorsal-ventral and rostral-caudal axes in the midbrain and hindbrain
5556 region and discovered early dorsal-ventral separation in the gut tube. Recently, Chen et
5557 al. applied high-resolution Stereo-seq to whole mouse embryos at the mid and
5558 late-gestation stage spanning from E9.5 to E16.5 and eventually constructed a mouse
5559 organogenesis spatiotemporal transcriptomic atlas (MOSTA) (Chen et al., 2022).

5560 Moving beyond early embryonic development in mice, many researchers have
5561 taken advantage of spatial transcriptomics to explore the spatially dependent
5562 mechanisms driving the development of organs or tissues in humans. For example,
5563 Crosse et al. utilized the LCM-based RNA sequencing to enable spatially resolved
5564 profiling of the developing hematopoietic stem cell (HSC) niche in human embryos at
5565 Carnegie stage (CS)16-CS17 (i.e., 39-41 post-conception days) (Crosse et al., 2020).
5566 They analyzed the dorsoventral polarized signaling in the aorta and identified ventrally
5567 secreted endothelin as an important secreted regulator of early human HSC
5568 development. In the study of the developing human heart, Asp et al. characterized
5569 different anatomical regions of human hearts at three developmental stages (4.5-5, 6.5
5570 and 9 post-conception weeks) using the ST technology (Asp et al., 2019). With the

5571 integration of scRNA-seq and ISS, a comprehensive spatial map was created, providing
 5572 detailed information about cell subtype localization during human cardiogenesis.
 5573 Similar strategies were applied to the developmental study of the human intestine
 5574 ranging from 8 to 22 PCW (Fawkner-Corbett et al., 2021). In addition to generating a
 5575 spatiotemporal atlas of human intestinal development, they also revealed how
 5576 morphogen gradients direct cellular differentiation. Spatial transcriptomics has also
 5577 been applied to the study of cell-type atlas and homeostasis maintenance in healthy
 5578 tissues of adults, which could serve as a reference to be compared with diseased tissues.
 5579 Shen et al. (Shen et al., 2023) applied the Stereo-seq technology to draw an ST atlas
 5580 of the human gingiva. By identifying periodontitis-relevant effector cells, genes and
 5581 pathways, the ST results may aid in the development of new therapeutic strategies for
 5582 periodontitis. By combining scRNA-seq, snRNA-seq, and 10x Visium ST, Madissoon
 5583 et al. created a spatial multi-omics atlas of the human lung and airway, which comprises
 5584 various novel and known cell types (Madissoon et al., 2021). The spatial lung atlas also
 5585 revealed specific tissue microenvironments, such as the gland-associated lymphoid
 5586 niche (GALN), which may play a role in preventing respiratory infections. In another
 5587 study of the human uterus, Garcia-Alonso et al. also applied multi-omics technologies
 5588 to construct a comprehensive cellular map of human endometrium, characterizing the
 5589 spatiotemporal dynamics across the menstrual cycle (Garcia-Alonso et al., 2021). In
 5590 particular, further spatial interaction analyses revealed the role of NOTCH and WNT
 5591 signaling pathways in shaping the differentiation of ciliated and secretory cell lineages.
 5592 With the accumulation of ST data, it is foreseeable that in the near future, integration of
 5593 multi-source tissue maps will lead to the establishment of a comprehensive reference
 5594 spatial atlas of the entire human body.

5595

5596 *Neuroscience*

5597 The explicit layered structures and distinct anatomical regions make the brain an
 5598 appropriate material to validate newly developed spatial transcriptomics technologies.
 5599 In return, these ST techniques significantly enhance our understanding of the spatial
 5600 architecture of brains. Many efforts have been devoted to building reference maps of
 5601 the brain. Due to the limited size of fields of view and intensive-labor nature of early
 5602 imaging-based ST techniques, most of the studies focused on specific subregions in the
 5603 mouse brain. For example, Codeluppi et al. developed osmFISH and employed the
 5604 methodology to define the spatial cellular organization of the somatosensory cortex,
 5605 covering only 33 targeted marker genes and around 5,000 cells (Codeluppi et al., 2018).
 5606 During the same time, Mottiff et al. generated a spatial molecular map of neurons in the
 5607 hypothalamic preoptic region by coupling MERFISH with scRNA-seq. Similarly, other
 5608 subregions of the brain, such as the visual cortex (Wang et al., 2018), the primary motor
 5609 cortex (Zhang et al., 2021a), the hippocampus (Shah et al., 2016; Alon et al., 2021), and
 5610 the cerebellum (Kebschull et al., 2020), have been profiled by different imaging-based
 5611 ST techniques to establish detailed spatial cellular organization maps.

5612 Thanks to the development of high-throughput barcoding-based ST techniques, a
 5613 molecular atlas of the whole adult mouse brain was established by Ortiz et al. (Ortiz et
 5614 al., 2020). They utilized the ST technology to profile spatial gene expression of 75

5615 adjacent coronal sections collected from one brain hemisphere along the
 5616 anteroposterior axis. Through alignment with the Allen mouse brain atlas (ABA), they
 5617 constructed a complete brain atlas, offering 3D tissue coordinates and detailed ABA
 5618 neuroanatomical definitions. More importantly, they also defined new area- and
 5619 layer-specific subregions in the molecular atlas by unsupervised classification.
 5620 Whether the entire brain or specific subregions are profiled, these atlases together will
 5621 be of great value to experimental neuroscience, ultimately extending our knowledge
 5622 about the structure-functional relationships of the brain.

5623 In addition to revealing the spatial organization of cell types in normal brains,
 5624 spatial transcriptomics can be extended to the study of neurodegenerative or psychiatric
 5625 diseases, uncovering spatially relevant mechanisms of dysfunction or dysregulation in
 5626 the nervous system. For example, Chen et al. combined the ST technology with ISS to
 5627 capture the transcriptional changes in the vicinity of amyloid plaques in Alzheimer's
 5628 disease (AD) (Chen et al., 2020b). In particular, they identified two gene co-expression
 5629 networks that might be responsive to amyloid plaque deposition in AD. In a study of
 5630 amyotrophic lateral sclerosis (ALS), Maniatis et al. employed the ST technology to
 5631 characterize the spatiotemporal dynamics over the progress of the disease by utilizing
 5632 murine models of ALS at different stages (Maniatis et al., 2019). Combining with
 5633 postmortem tissues from ALS patients, they discerned shared spatial patterns of
 5634 perturbations in transcriptional pathways associated with ALS pathology.

5635 As ST technologies continue to improve in resolution and detection efficiency, we
 5636 anticipate the establishment of more detailed and comprehensive atlases of the nervous
 5637 system. These resources will undoubtedly be invaluable for exploring the
 5638 structure-function relationships of circuits and behaviors.

5639

5640 *Tumor microenvironment*

5641 Although single-cell transcriptomics has shed light on the cell-type compositions and
 5642 their functions in the intricate tumor microenvironment (TME), it remains unexplored
 5643 how these cells are spatially organized to control or promote tumor progression. Spatial
 5644 transcriptomics makes it possible to study different cell populations and signaling
 5645 pathways with the spatial context preserved. Generally, tumor microenvironments
 5646 might include tumor cells, stromal cells, and immune cells. Initial research efforts often
 5647 concentrate on is the interior heterogeneities of tumor regions. In a single-cell and
 5648 spatial atlas study of human breast cancer (BRCA), Wu et al. derived seven gene
 5649 modules from scRNA-seq to describe the intratumor transcriptional heterogeneity (Wu
 5650 et al., 2021d). The enrichment analysis revealed two gene modules mutually exclusive
 5651 in the tumor regions, which were related to the epithelial-mesenchymal transition
 5652 (EMT) and proliferation states, respectively. In another study of primary liver cancer,
 5653 five cancer stem cell (CSC) populations were defined, which showed different
 5654 distribution patterns in different regions, including the leading edge, tumor, and
 5655 high-grade portal vein tumor thrombosis (Wu et al., 2021b). Of note, the fraction of
 5656 PROM1⁺ CSCs was higher in portal vein tumor thrombosis than in the tumor region,
 5657 potentially exerting a crucial role in the tumor progression.

5658 Centered on the tumor region, the relative spatial distribution of immune or

stromal cell types could be revealed by spatial transcriptomics. In the study of human squamous cell carcinoma (SCC), Ji et al. discovered that B cells were infiltrated in the tumor, while regulatory T cells, macrophages, and fibroblasts were abundant at the tumor-stromal border. Conversely, CD8 T cells were notably excluded from the tumor (Ji et al., 2020a). Similarly, different cell subtypes or states also reveal different spatial patterns. Wu et al. identified both inflammatory-like cancer-associated fibroblasts (iCAFs) and myofibroblast-like CAFs (mCAFs) in the TME of breast cancer, yet the two subtypes exhibited markedly distinct spatial distributions (Wu et al., 2021d). mCAFs were found to be enriched in invasive cancer regions, while iCAFs were dispersedly distributed across invasive cancer, stroma, and lymphocytes-aggregate regions. Some studies are interested in the molecular and cell-type patterns in the tumor-stromal border, namely the invasive fonts of tumors (Hunter et al., 2021). Wu et al. characterized the dynamics of cell-type abundance across the invasive fonts and found an immune suppressive microenvironment in the area near the borderline (Wu et al., 2021a).

The spatial analysis could also recognize some patterned structures and characterize them in the tumor microenvironment. In the abovementioned study of liver cancer (Wu et al., 2021b), unsupervised clustering of ST spots revealed a cluster characterized by high expression of tertiary lymphoid structures (TLS)-related genes, such as CXCL13, CCL19, CCL21, LTF, and LTB. The pathological examination validated the presence of TLSs. Then, Wu et al. defined a TLS-50 signature to locate TLSs in other tissue sections, which was also found to be associated with more favorable prognosis in HCC patients in TCGA. Similarly, Andersson et al. also identified TLSs in HER2-positive breast cancer (Andersson et al., 2021b). To further investigate how TLSs influence the response to immunotherapy in cancer, Meylan et al. used spatial transcriptomics to examine the nature of B cell responses within TLS in renal cell carcinoma (RCC) (Meylan et al., 2022). They discovered that TLSs could generate and propagate anti-tumor antibody-producing plasma cells, which is associated with response to immunotherapy.

Cellular communications are known to play important roles in the immune surveillance or escape of tumors, as well as tumor progression. With the spatial distribution of cell types revealed by cell-type deconvolution or cell mapping analysis, cell-type proximity or colocalization patterns could also be recognized. Moncada et al. identified the colocalization of inflammatory fibroblasts and stress-response cancer cells by mapping scRNA-seq-defined cell types to ST of pancreatic ductal adenocarcinomas (Moncada et al., 2020). Similarly, with the integration of scRNA-seq and ST in SCC, a fibrovascular niche was found to surround a tumor-specific keratinocyte population (Ji et al., 2020a). Further interaction analysis revealed the colocalization might be mediated by multiple ligand-receptor pairs. In another study of colorectal cancer, spatial transcriptomics and immunofluorescent staining demonstrated the co-existence of FAP⁺ fibroblasts and SPP1⁺ macrophages, which was associated with poor patient survival (Qi et al., 2022).

With the development of spatial multi-omics techniques, additional facets such as cell crosstalk and metabolic states, will be characterized to gain more insights into the

5703 complexity of the tumor microenvironment. Understanding the tumor
 5704 microenvironment will facilitate the identification of therapeutic targets and the design
 5705 of anti-tumor drugs.

5706

5707 *Summary*

5708 This chapter provides a comprehensive overview of current advancements in spatial
 5709 transcriptomics, encompassing technical innovations, computational methods, and
 5710 diverse applications. Spatial transcriptomics has revolutionized our understanding of
 5711 tissue organization and cellular heterogeneity, enabling high-resolution visualization of
 5712 gene expression patterns within intact tissues. The development of computational
 5713 methods has facilitated the integration and interpretation of spatial transcriptomics data,
 5714 unveiling spatial regulatory mechanisms and novel molecular interactions. Spatial
 5715 transcriptomics has been successfully applied in various fields, including
 5716 developmental biology, neuroscience, cancer research, and immunology, with the
 5717 potential to accelerate biomarker discovery and personalized medicine approaches.
 5718 Spatial transcriptomics represents a transformative approach and will continue to be
 5719 refined to reshape our understanding of complex biological systems. We anticipate it
 5720 will offer profound insights into tissue homeostasis and disease mechanisms.

5721

5722 **Chapter 8 Single-cell CRISPR screening technology**

5723

5724 The clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 system
 5725 is a revolutionary approach to edit the mammalian genome (Cong et al., 2013; Mali et
 5726 al., 2013). With the development of lentiviral delivery methods, CRISPR screening
 5727 technology emerged and has enabled genome-wide knockout in a cost-effective
 5728 manner (Koike-Yusa et al., 2014; Shalem et al., 2014; Wang et al., 2014a). However,
 5729 CRISPR screening can only analyze genes with very distinct phenotypes, such as
 5730 those that significantly affect cell growth or those that can be detected directly with
 5731 antibodies or fluorescent proteins.

5732 In 2016, a new technique, called single-cell CRISPR screening (scCRISPR-seq),
 5733 was developed that coupled CRISPR perturbations and single-cell sequencing to
 5734 enable pooled genetic screens at large-scale single-cell resolution (Adamson et al.,
 5735 2016; Dixit et al., 2016; Jaitin et al., 2016). The key technical innovation of
 5736 scCRISPR-seq is the creative design of the lentiviral vector, called the Perturb-seq
 5737 vector, to allow the identification of sgRNA in each cell from sequencing (Figure
 5738 16A). scCRISPR-seq can facilitate high-throughput functional dissection of complex
 5739 regulatory mechanisms and heterogeneous cell populations.

5740 In this chapter, we will comprehensively review scCRISPR-seq in four distinct
 5741 parts. Firstly, we will introduce representative technologies within each category of
 5742 scCRISPR-seq. Secondly, we will delve into the primary tools that have been
 5743 specifically developed for the analysis of scCRISPR-seq data. Thirdly, we will
 5744 explore notable applications of scCRISPR-seq. Finally, we will draw conclusions and
 5745 engage in a discussion of the limitations and future trends associated with
 5746 scCRISPR-seq.

5747

5748 *The category of scCRISPR-seq platforms*

5749 Currently, numerous alternative scCRISPR-seq platforms have emerged (Table 5).
 5750 Based on the integrated omics approach of scCRISPR-seq, these platforms can be
 5751 classified into three primary categories: transcriptome-based scCRISPR-seq,
 5752 epigenome-based scCRISPR-seq, and multimodal scCRISPR-seq.

5753

5754 *Transcriptome-based scCRISPR-seq*

5755 The main scCRISPR-seq platforms are transcriptome-based applications that combine
 5756 CRISPR screens with single-cell RNA-seq. For transcriptome-based scCRISPR-seq,
 5757 the Perturb-seq vector is generally composed of single-guide RNA (sgRNA), cell
 5758 barcode (CBC), gene barcode (GBC), and unique molecular identifier (UMI), such as
 5759 Perturb-seq (Adamson et al., 2016; Dixit et al., 2016) and CRISP-seq (Jaitin et al.,
 5760 2016). In the Perturb-seq vector, sgRNA is used to direct Cas9 nucleases to induce
 5761 double-strand breaks at targeted genomic regions and CBC is used to tag each cell,
 5762 while GBC is used to tag each sgRNA and UMI is used to tag each transcript.
 5763 Perturb-seq and CRISP-seq are the first scCRISPR-seq platforms to be developed.
 5764 These approaches involved complex construction of the Perturb-seq vector, including
 5765 complex cloning strategy and sometimes the decoupling of the gRNA spacer and its
 5766 barcode, which limits their versatility. The CROP-seq (Datlinger et al., 2017)
 5767 optimizes the design of the Perturb-seq vector to allow the detection of sgRNA
 5768 induced in each cell coupled with mRNA by adding Poly-A tail to the Perturb-seq
 5769 vector, which greatly reduces the complexity and cost of scCRISPR-seq. However,
 5770 Hill et al. (Hill et al., 2018) demonstrated that the lentivirus swap rate of existing
 5771 studies was only about 50% because of the Perturb-seq vector designs of these studies.
 5772 Thus, they optimized CROP-seq vector designs by serving the guide RNA as the
 5773 barcode to improve the swap rate to 94%. Due to constraints on Perturb-seq vector
 5774 design, each lentiviral vector of Perturb-seq and CRISP-seq can only deliver a single
 5775 encoded sgRNA to cells, and CROP-seq enabled the delivery of paired sgRNAs to
 5776 cells. That is, they are all incompatible with the delivery of multiple sgRNAs. To
 5777 solve this problem, Replogle et al. (Replogle et al., 2020) designed direct-capture
 5778 Perturb-seq, in which expressed sgRNAs are sequenced alongside single-cell
 5779 transcriptomes and enable the delivery of multiple sgRNAs. Direct-capture
 5780 Perturb-seq is particularly valuable for the mechanistic dissection of genetic
 5781 interaction. It further reduces the cost of Perturb-seq experiments. Direct-seq (Song et
 5782 al., 2020) has similar functions as direct-capture Perturb-seq that enables CRISPR
 5783 perturbation and its transcriptional readouts profiled together and supports the
 5784 delivery of multiple sgRNAs. In 2022, the genome-scale Perturb-seq method was
 5785 introduced by Replogle et al., enabling unbiased and comprehensive profiling of
 5786 genome-scale genetic perturbations affecting 9867 genes (Replogle et al., 2022). This
 5787 breakthrough facilitated systematic gene function assignment and the exploration of
 5788 complex cellular phenotypes. More recently, Li et al. (Li et al., 2023) developed the
 5789 CRISPR-human organoids-single-cell RNA sequencing (CHOOSE) system. This
 5790 innovative system enables genetic disruption and single-cell transcriptomics for

5791 pooled loss-of-function screening in mosaic organoids.

5792 However, all the aforementioned scCRISPR-seq platforms are limited to in vitro
 5793 applications. In contrast, in vivo assays are more attractive due to their closer
 5794 resemblance to real organic conditions. Therefore, Jin et al. developed in vivo
 5795 Perturb-seq (Jin et al., 2020b), a variation of the Perturb-seq protocol that involves
 5796 pooled perturbations conducted in vivo. Furthermore, PoKI-seq (Roth et al., 2020) has
 5797 demonstrated the feasibility of in vivo investigation of the immunological response of
 5798 reprogrammed T cells to solid tumors. Recently, Santinha et al. (Santinha et al., 2023)
 5799 developed adeno associated virus (AAV)-mediated direct in vivo single-cell CRISPR
 5800 screening, termed AAV-Perturb-seq, a tunable and broadly applicable method for
 5801 transcriptional linkage analysis and phenotyping of genetic perturbations in vivo.
 5802

5803 *Epigenome-based scCRISPR-seq*

5804 In addition to transcriptome applications, there are also epigenetic-based
 5805 scCRISPR-seq platforms. In 2019, Rubin et al. (Rubin et al., 2019) developed
 5806 Perturb-ATAC, a method that combines CRISPR interference or knockout with
 5807 chromatin accessibility profiling in single cells based on simultaneous detection of
 5808 CRISPR guide RNAs and open chromatin sites by assay of transposase-accessible
 5809 chromatin with sequencing (ATAC-seq). They applied this method to determine the
 5810 roles of a diverse set of trans-regulatory factors, including TFs, chromatin modifiers,
 5811 and human and viral ncRNAs, which may be useful for dissecting loci where both
 5812 cis-regulatory elements and ncRNA transcripts have been shown to have effects on
 5813 gene expression (Engreitz et al., 2016; Cho et al., 2018b; Rubin et al., 2019).
 5814 Perturb-ATAC expands scCRISPR-seq research into the epigenome field, making
 5815 scCRISPR-seq more powerful and broader applications. However, Perturb-ATAC is
 5816 constrained by high costs and low throughput. In response to this limitation,
 5817 Spear-ATAC (Pierce et al., 2021) was developed to achieve significantly higher cell
 5818 throughput and a substantial cost reduction, offering a more practical alternative.
 5819 Additionally, CRISPR-sciATAC (Liscovitch-Brauer et al., 2021) demonstrated
 5820 similar cell throughput and cost to Spear-ATAC. However, it exhibited limited
 5821 sensitivity to subtle changes in chromatin accessibility.
 5822

5823 *Multimodal scCRISPR-seq*

5824 Multimodal single-cell assays provide high-resolution snapshots of heterogeneous cell
 5825 populations, but the scCRISPR-seq platforms above are all limited to one modality,
 5826 such as transcriptome or epigenome. Thus, to apply the technique to multi-omics
 5827 simultaneously, multimodal scCRISPR-seq was developed. Xie et al. (Xie et al., 2017)
 5828 developed mosaic single-cell analysis by indexed CRISPR sequencing (Mosaic-seq)
 5829 to perturb enhancers and jointly measure each cell's transcriptome and its induced
 5830 sgRNA. Mosaic-seq provides a novel tool to interrogate the functions of noncoding
 5831 genes in a perturbation-based manner. In addition, Mimitou et al. (Mimitou et al.,
 5832 2019) developed expanded CRISPR-compatible cellular indexing of transcriptomes
 5833 and epitopes by sequencing (ECCITE-seq), which allowed simultaneous detection of
 5834 transcriptomes, proteins, clonotypes, and CRISPR perturbations from every single cell.

By constructing a 49-marker panel of ECCITE-seq antibodies to profile human peripheral blood mononuclear cells (PBMCs), they recovered many important results (Stoeckius et al., 2017; Fanok et al., 2018), demonstrating the power of ECCITE-seq to combine immunophenotype, clonotype, and transcriptome information. Spatial transcriptomics is able to characterize gene expression profiles while retaining information about the spatial tissue context, which provides new insights into different fields of biology, such as neuroscience, developmental biology, and cancer research (Moses and Pachter, 2022) (see the spatial section below). Recently, a new multimodal scCRISPR-seq called Perturb-map (Dhainaut et al., 2022) was developed to enable multimodal phenotyping of CRISPR screens *in situ* by imaging and spatial transcriptomics. Perturb-map is based on a protein bar code (Pro-Code) system that uses triplet combinations of a few linear epitopes to create a higher-order set of unique bar codes (Wroblewska et al., 2018). These unique bar codes can mark cells expressing different CRISPR gRNAs. It should be noted that Perturb-map is the only scCRISPR-seq platform that enables *in vivo* CRISPR screens combined with spatial transcriptome, which is particularly suitable for the identification of genetic determinants of tumor composition, organization, and immunity. Dhainaut et al. (Dhainaut et al., 2022) applied Perturb-map to the study of the tumor microenvironment (TME). They knocked out 35 genes in a mouse model of lung cancer and found that knockout of Tgfbr2 can promote TME remodeling and immune exclusion.

The tools to analyze scCRISPR-seq data

scCRISPR-seq data contain rich perturbation information, which is a natural advantage to exploring the association between genotype and phenotype at a single cell level. For example, by applying Perturb-seq to the K562 cell line, Adamson et al. (Adamson et al., 2016) have shown that perturbation of PERK has a greater impact on the unfolded protein response than ATF6 and IRE1 α . Datlinger et al. (Datlinger et al., 2017) perturbed 23 transcription factors in the Jurkat cell line under the condition of T cell receptor (TCR) activation with CROP-seq and found that knockouts of LCK, ZAP70, and LAT have a strong negative effect on TCR activation signaling.

However, the analysis of scCRISPR-seq data is a major challenge due to its inherent noise. Thus, several bioinformatic tools have been developed to help analyze scCRISPR-seq data (Table S12). Generally, these scCRISPR-seq data analysis tools focus on three parts (Figure 16B): (1) Data preprocessing, including quality control, normalization, and differentially expressed genes detection, such as MIMOSCA (Dixit et al., 2016), MUSIC (Duan et al., 2019), and SCREE (Wei et al., 2023). (2) Data denoise, including single-cell imputation, escaping cells filtering, and compound factors decomposing, such as MUSIC, mixscape (Papalexis et al., 2021), and SCREE (Wei et al., 2023). (3) Functional analysis, including prioritizing the impact of each perturbation, identifying the function of each perturbation, inferring regulatory network and gene interaction, such as MUSIC, Normalisr (Wang, 2021), scMAGeCK (Yang et al., 2020), Pando (Fleck et al., 2023) and GEARS (Roohani et al., 2023). Specifically, LRICA is proposed to decode the driver signal/component of the data by

5879 low-rank matrix factorization. MIMOSCA is a computational framework for
 5880 calculating the relationship between sgRNA and each gene. LRICA and MIMOSCA
 5881 were developed as prototypes without executable and user-friendly implementations.
 5882 Thus, Duan et al. (Duan et al., 2019) developed MUSIC, a general computational
 5883 framework to evaluate the impact of each perturbation with topic modeling (Blei,
 5884 2007), which was originally presented in the machine learning and natural language
 5885 processing community for latent topic discovery in a particular set of documents.
 5886 MUSIC links genotype to phenotype with tolerance to substantial noise and analyzes
 5887 scCRISPR-seq data from three perspectives, i.e., prioritizing the gene perturbation
 5888 effect as an overall perturbation effect, in a functional topic-specific manner, and
 5889 quantifying correlations between different perturbations. scMAGECK is also a
 5890 framework for analyzing scCRISPR-seq data, which is extended from MAGeCK (Li
 5891 et al., 2014). scMAGECK includes two modules, scMAGECK-RRA and
 5892 scMAGECK-LR, where scMAGECK-RRA is used to identify significantly enriched
 5893 sgRNAs by the negative binomial distribution, and scMAGECK-LR is used to assess
 5894 affected genes by linear regression. scMAGECK showed a good control of false
 5895 positives and better sensitivity than other methods. In addition to the general
 5896 computational framework of scCRISPR-seq, some tools focus on data denoising. For
 5897 example, SCEPTRE was developed for scCRISPR-seq data calibration using
 5898 conditional randomization testing. SCEPTRE demonstrated good calibration and
 5899 sensitivity to scCRISPR-seq data, yielding hundreds of new regulatory relationships
 5900 supported by orthogonal biological evidence. mixscape was developed to improve the
 5901 signal-to-noise ratio of scCRISPR-seq data by filtering escaping cells (cells induced
 5902 sgRNA, but did not exhibit perturbation effect) by mixed discriminant analysis.
 5903 Normalisr is developed to reconstruct gene regulatory network for scCRISPR-seq data.
 5904 Wang et al. (Wang et al., 2022g) emphasized the significance of identifying clone
 5905 cells, as they can lead to false positives in scCRISPR-seq data. SCREE serves as a
 5906 comprehensive pipeline for scCRISPR-seq data analysis. In contrast to the previously
 5907 mentioned approaches, which primarily concentrated on data denoising and mining in
 5908 scCRISPR-seq, GEARS was specifically designed to predict transcriptional responses
 5909 to both single and multigene perturbations. These methodologies have substantially
 5910 enhanced the analysis of scCRISPR-seq data.

5911

5912 *Applications of scCRISPR-seq*

5913 scCRISPR-seq is widely applied in various fields due to its powerful capabilities,
 5914 including linking genotype to phenotype, dissecting genetic regulations, and
 5915 investigating genetic mechanisms in specific diseases, such as tumor and autism.

5916

5917 *Linking genotype to phenotype*

5918 Compared to traditional CRISPR screening, which can only identify genes with very
 5919 distinct phenotypes, scCRISPR-seq has the ability to uncover the functions of any
 5920 genes. Therefore, scCRISPR-seq is naturally suited for linking genotype to phenotype
 5921 on a large scale. For example, Jaitin et al. (Jaitin et al., 2016) revealed the effect of 22
 5922 TFs on the regulation of antiviral, inflammatory, or developmental processes in

5923 Lipopolysaccharide (LPS) stimulated bone marrow cells (BMCs) by CRISP-seq.
 5924 Adamson et al. (Adamson et al., 2016) analyzed systematically the effect of 83
 5925 unfolded protein response (UPR) related genes in K562 cells by Perturb-seq. In
 5926 addition, genome-scale Perturb-seq (Replogle et al., 2022) offers unbiased,
 5927 comprehensive profiling of genetic perturbations (9867 genes), facilitating
 5928 systematical dissection of relationships between genes related to gene translation and
 5929 ribosome biogenesis.

5930

5931 *Dissecting genetic regulations*

5932 scCRISPR-seq is also used to dissect complex relationships between genomic
 5933 elements, including coding genes, transcription factors, chromatin regulators,
 5934 enhancers, and other non-coding elements. For example, Adamson et al. (Adamson et
 5935 al., 2016) discovered the crosstalk between three UPR sensor genes (ATF6, PERK,
 5936 and IRE1) using Perturb-seq. CROP-seq perturbed TFs regulating TCR activation in
 5937 Jurkat cells upon LPS stimulation and uncovered the relationship between TFs. In
 5938 addition, scCRISPR-seq for enhancer perturbation, such as Mosaic-seq (Xie et al.,
 5939 2017), could discover novel enhancer-gene pairs. In addition, scCRISPR-seq coupled
 5940 with scATAC-seq, such as Perturb-ATAC, Spear-ATAC, and CRISPR-sciATAC
 5941 could reveal epigenetic landscape remodelers in human B lymphocytes and leukemia
 5942 cells (Rubin et al., 2019; Liscovitch-Brauer et al., 2021; Pierce et al., 2021).

5943

5944 *Investigating genetic mechanisms*

5945 Several *in vivo* scCRISPR-seq platforms are available, enabling studies of genetic
 5946 mechanisms in specific diseases such as tumors and autism. For instance, Perturb-map
 5947 (Dhainaut et al., 2022) facilitates the identification of genetic determinants related to
 5948 tumor composition, organization, and immunity. Using Perturb-map, Dhainaut et al.
 5949 discovered that the knockout of tgfbr2 in lung cancer cells promotes tumor
 5950 microenvironment remodeling and immune exclusion. Roth et al. conducted a screen
 5951 for chimeric antigen receptors that enhance T cell anti-tumor functions, improving
 5952 tumor infiltration and cell killing rates under immunosuppressive conditions in
 5953 melanoma using PoKI-seq (Roth et al., 2020). Furthermore, Jin et al. evaluated 35 *de*
 5954 *novo* loss-of-function risk genes associated with autism spectrum
 5955 disorder/neurodevelopmental delay (ASD/ND) using *in vivo* Perturb-seq. They
 5956 identified cell type-specific and evolutionarily conserved gene modules from both
 5957 neuronal and glial cell classes (Jin et al., 2020b). Li et al. also focused on these
 5958 high-risk autism spectrum disorder genes, and they uncovered their effects on cell fate
 5959 determination in mosaic organoids with CHOOSE system (Li et al., 2023). Recently,
 5960 Santinha et al. (Santinha et al., 2023) employed AAV-Perturb-seq to systematically
 5961 analyze the phenotypic landscape associated with 22q11.2 deletion syndrome genes in
 5962 the prefrontal cortex of adult mouse brains. They identified three 22q11.2-linked
 5963 genes actively involved in both established and previously unrecognized pathways
 5964 governing neuronal functions *in vivo*.

5965

5966 *Summary*

5967 In this chapter, we have presented a comprehensive review of scCRISPR-seq, divided
 5968 into three distinct parts, which include the categories of scCRISPR-seq, tools for the
 5969 analysis of scCRISPR-seq data, and notable applications of scCRISPR-seq.

5970 scCRISPR-seq has been a powerful approach for functional genomics research
 5971 (Bock, 2022). In this section, we have categorized scCRISPR-seq into three primary
 5972 categories based on its integrated omics approach: transcriptome-based
 5973 scCRISPR-seq, epigenome-based scCRISPR-seq, and multimodal scCRISPR-seq.
 5974 Given the inherent noise in scCRISPR-seq data, a multitude of bioinformatic tools
 5975 have been developed to aid in its analysis, resulting in significant improvements.

5976 The versatility of scCRISPR-seq has led to its widespread application across
 5977 various fields, offering potent capabilities such as connecting genotype to phenotype,
 5978 dissecting genetic regulations, and exploring genetic mechanisms in specific diseases
 5979 like tumors and autism. Nevertheless, before its broader adoption in biological
 5980 research, three key aspects require attention: (1) Reducing complexity and cost:
 5981 efforts should be made to further streamline and reduce the complexity and cost of
 5982 scCRISPR-seq experiments. This will enhance scalability and accessibility, allowing
 5983 more laboratories to leverage this technology. (2) Expanding applicability to complex
 5984 tissues and *in vivo* settings: while current scCRISPR-seq platforms primarily target
 5985 cell lines, there is a pressing need to develop more robust scCRISPR-seq platforms
 5986 that can be applied to more complex tissues, including organoids, and ideally *in vivo*
 5987 settings. This expansion will enable a broader range of biological investigations. (3)
 5988 Noise reduction techniques: As the number of scCRISPR-seq platforms grows, it
 5989 becomes crucial to develop more powerful methods for deciphering the inherent noise
 5990 in scCRISPR-seq data. These methods will contribute to the reliability and
 5991 interpretability of scCRISPR-seq results, further enhancing their utility in diverse
 5992 research contexts.

5993

5994 Epilogue

5995

5996 scRNA-seq technology has attracted widespread attention from many scientists
 5997 around the world because it has the advantage of providing an unprecedented method
 5998 to study cell heterogeneity at the single-cell level. A mere 14 years have elapsed since
 5999 the establishment of a new era in scRNA-seq research, which was preceded by the
 6000 initial conceptual and technical breakthrough achieved by Tang et al. in 2009 (Tang et
 6001 al., 2009). The field of scRNA-seq research is currently experiencing a surge in
 6002 studies, driven by the continuous development of sequencing technology and
 6003 bioinformatics.. The maturity of scRNA-seq technology has greatly facilitated
 6004 advancements in other single-cell omics studies. At present, single-cell omics
 6005 detection has been extended to the genome (Dey et al., 2015), epigenome (Muto et al.,
 6006 2021), spatial transcriptomics (Chen et al., 2022), proteome (Peterson et al., 2017;
 6007 Specht et al., 2021), metabolome (Shrestha, 2020) and other multiomics levels
 6008 (Angermueller et al., 2016), providing a more comprehensive, refined and complete
 6009 analysis strategy for single-cell level research. In this review, we summarize the
 6010 state-of-the-art developments in single-cell omics technologies, data analyses, and

their applications, outlining the landscape of the single-cell sequencing field across multiple layers.

In Chapter 1, we provide a comprehensive overview of the currently available scRNA-seq technologies, experimental methodologies, data analysis procedures, and their applications within the biomedical field. Initially, single-cell sequencing was performed by isolating single cells and independently constructing a sequencing library. These single-cell sequencing technologies can only detect a small number of cells (tens to hundreds), such as the Tang method, STRT-seq, and SMART-seq (Tang et al., 2009; Islam et al., 2012; Ramsköld et al., 2012). However, with the in-depth study of sequencing technology, single-cell identification based on barcode tags has emerged, and with the emergence of new single-cell separation technologies based on microdroplets or microwells such as Drop-Seq and Cyto-Seq (Fan et al., 2015a; Macosko et al., 2015), and single-cell transcriptome sequencing has entered the era of high-throughput. The sequencing cost has been dramatically reduced, while automation and throughput have been significantly increased. scRNA-seq technology solves the problem of cell heterogeneity, opens new avenues for personalized treatment of clinical diseases, especially tumors, and promotes the development of precision medicine. However, scRNA-seq has limitations of low capture efficiency and high dropouts due to the low amount of starting material. Compared to bulk RNA-seq, scRNA-seq produces noisier and more variable data. Although, researchers have designed a variety of tools to conduct diverse scRNA-seq data analyses, the technical noise and biological variation (e.g., stochastic transcription) still pose huge challenges for computational analysis of scRNA-seq data (Chen et al., 2019a). Therefore, data analysis methods still need to be further optimized and improved.

Compared with the increasingly mature scRNA-seq technology, the other single-cell omics technologies are still budding. In Chapters 2, 3, 4, and 5, we focus on the state-of-the-art tools, computational methods, and applications for single-cell genome, epigenome, proteomic, and metabolomics sequencing over the past ten years. ScWGS has revolutionized our understanding of genetic variation and its impact on human health and disease. The rapid development of its has accelerated genomic research, enabled personalized medicine, and provided valuable insights into the genetic basis of diseases and human diversity. Cells exhibit extensive heterogeneity in terms of chromatin accessibility, nucleosome positioning, histone modifications, and DNA methylation. Mapping this epigenomic information in single-cell samples is very important for developmental biology, cancer research, and so on. Advances in single-cell epigenomic sequencing approaches are enabling high-resolution mapping of chromatin states in single cells. However, nowadays single-cell epigenomic techniques suffer from data loss. As a result, even though individual cell epigenomic data sets are powerful resources for clustering analyses and for revealing cellular heterogeneity based on the collection of a great number of target sites, they have only very limited ability to provide information on single target sites (Carter and Zhao, 2021). Therefore, improving the coverage of chromatin target sites in various individual cell epigenomic assays will be required in future studies, which will contribute to understanding cell heterogeneity at a whole cell level and single specific

site. Single-cell proteomics is in the early stage of explosive development due to its complex constituents, low abundance, wide dynamic range, and lack of amplified ability. Just in 2019, analysis of the proteome from single cells was described as a “dream”, but today there have been several promising tools developed (Marx, 2019). We believe that with the optimization of accessibility and the further improvement of throughput, the truly large-scale applications of single-cell proteomics in scientific and clinical research, such as organ maps, drug screening, and precise disease classification, are within reach. Single-cell metabolomics is used to identify the composition of metabolites in a single cell, measure their abundance, and study their dynamic changes. Meanwhile, the metabolome represents the downstream products of the genome, transcriptome, and proteome, and provides a more immediate and dynamic snapshot of the functionality (Shrestha, 2020). Overall, single-cell omics technologies are still in the budding stage, it is going to continue to flourish.

A single cell serves as the fundamental unit of life. Multi-omics analysis of a single cell can offer profound insights into the cell’s phenotype, disease state, and environmental impacts. In Chapters 6, 7, and 8, we comprehensively summarize the integrated analysis of multi-omics, the combined application of scRNA-seq and CRISPR screening, and spatial transcriptome. In intricate biological processes, such as tumorigenesis and aging, heterogeneity occurs on different levels, including the genome, transcriptome, proteome, and epigenome. If only one component is analyzed from a single cell at a time, only the local overview of the gene regulatory network can be detected, while the complex global situation cannot be accurately predicted. In this situation, multi-omics technology highlights its unique advantages, which can provide a more complete map of the gene regulatory network in the study of complex tissues. For spatial transcriptomics, it has enabled the measurement of gene expression with spatial information preserved, which will be conducive to investigating intercellular relationships and discovering novel regulation mechanisms in the spatial context. In addition, spatial transcriptomics makes it possible to explore the spatial regulation mechanisms of cell fate determination and the architecture of tissue patterning. Compared with traditional CRISPR hybrid screening, the combination of scRNA-seq and CRISPR can not only screen thousands of gRNAs in a single experiment but also simultaneously captures perturbed full-transcriptome data for the clearest understanding of cell type specific gene function and pathway analysis. Therefore, the combination of these techniques enables a better and deeper understanding of key biological processes and mechanisms, which is an important direction for the development of single-cell technology in the future.

Nowadays, single-cell omics technologies have witnessed significant advancements in terms of both throughput and resolution. Moving forward, the primary trends in single-cell technology development are to improve the efficiency and throughput of single-cell sorting, enhance the sequencing coverage and sensitivity, realize high-throughput multi-omics studies, and develop more automated single-cell technology platforms, which will help reduce the cost and technical threshold of single-cell technology. The single-cell technology promises to be widely used in the

6098 field of scientific research and research transformation and will have a great
 6099 contribution to health monitoring, disease diagnosis, and treatment.
 6100

6101 **Compliance and ethics**

6102 *The authors declare that they have no conflict of interest.*

6104 **Acknowledgements**

6105 We would like to express our sincerest gratitude to the members of the expert panel of
 6106 the Program on Single- Cell Omics. This work was supported by the National Natural
 6107 Science Foundation of China (32130020, 32025009, 82030099, 30700397, 31970638,
 6108 61572361, 81973701, U23A20513), the National Key Research and Development
 6109 Program of China (2021YFF1201200, 2021YFF1200900), the Shanghai Municipal
 6110 Science and Technology Major Project (2017SHZDZX01), Beihang University &
 6111 Capital Medical University Plan (BHME-201904), the Special Fund of the Pediatric
 6112 Medical Coordinated Development Center of Beijing Hospitals Authority
 6113 (XTCX201809), the Cooperative Research Fund of the Affiliated Wuhu Hospital of
 6114 East China Normal University (40500-20104-222400), the Fundamental Research
 6115 Funds for the Central Universities (226-2024-00001), the Shanghai Municipal Science
 6116 and Technology Commission “Science and Technology Innovation Action
 6117 Plan” technical standard project (21DZ2201700), the Shanghai Municipal Science
 6118 and Technology Commission “Science and Technology Innovation Action Plan”
 6119 natural science foundation project (23ZR1435800), the Shanghai Natural Science
 6120 Foundation Program (17ZR1449400), the Shanghai Artificial Intelligence Technology
 6121 Standard Project (19DZ2200900), the Open Research Fund of Key Laboratory of
 6122 Advanced Theory and Application in Statistics and Data Science-MOE, ECNU, Key
 6123 Laboratory of MEA, Ministry of Education, ECNU.
 6124

6125 **References**

- 6126 A, M.A., Ibáñez-Solé, O., Inza, I., Izeta, A. and Araúzo-Bravo, M.J. (2022). Triku: a feature
 6127 selection method based on nearest neighbors for single-cell data. *Gigascience* 11, giac017.
 6128 Abdelfattah, N., Kumar, P., Wang, C., Leu, J.S., Flynn, W.F., Gao, R., Baskin, D.S., Pichumani, K.,
 6129 Ijare, O.B., Wood, S.L., et al. (2022). Single-cell analysis of human glioma and immune cells
 6130 identifies S100A4 as an immunotherapy target. *Nat Commun* 13, 767.
 6131 Abid, A., Zhang, M.J., Bagaria, V.K. and Zou, J. (2018). Exploring patterns enriched in a dataset
 6132 with contrastive principal component analysis. *Nat Commun* 9, 2134.
 6133 Abouleila, Y., Onidani, K., Ali, A., Shoji, H., Kawai, T., Lim, C.T., Kumar, V., Okaya, S., Kato, K.,
 6134 Hiyama, E., et al. (2019). Live single cell mass spectrometry reveals cancer-specific
 6135 metabolic profiles of circulating tumor cells. *Cancer Sci* 110, 697-706.
 6136 Achim, K., Pettit, J.B., Saraiva, L.R., Gavriouchkina, D., Larsson, T., Arendt, D. and Marioni, J.C.
 6137 (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat
 6138 Biotechnol* 33, 503-509.
 6139 Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert,
 6140 L.A., Horlbeck, M.A., Hein, M.Y., et al. (2016). A Multiplexed Single-Cell CRISPR
 6141 Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*

- 6142 167, 1867-1882.e1821.
- 6143 Adossa, N., Khan, S., Rytkonen, K.T. and Elo, L.L. (2021). Computational strategies for
6144 single-cell multi-omics integration. *Comput Struct Biotechnol J* 19, 2588-2596.
- 6145 Affinati, A.H., Sabatini, P.V., True, C., Tomlinson, A.J., Kirigiti, M., Lindsley, S.R., Li, C., Olson,
6146 D.P., Kievit, P., Myers, M.G., *et al.* (2021). Cross-species analysis defines the conservation of
6147 anatomically segregated VMH neuron populations. *Elife* 10, e69065.
- 6148 Ai, S., Xiong, H., Li, C.C., Luo, Y., Shi, Q., Liu, Y., Yu, X., Li, C. and He, A. (2019). Profiling
6149 chromatin states using single-cell itChIP-seq. *Nat Cell Biol* 21, 1164-1172.
- 6150 Aicher, T.P., Carroll, S., Raddi, G., Gierahn, T., Wadsworth, M.H., 2nd, Hughes, T.K., Love, C.
6151 and Shalek, A.K. (2019). Seq-Well: A Sample-Efficient, Portable Picowell Platform for
6152 Massively Parallel Single-Cell RNA Sequencing. *Methods Mol Biol* 1979, 111-132.
- 6153 Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C.
6154 and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina
6155 sequencing libraries. *Genome Biol* 12, R18.
- 6156 Akar-Ghibril, N. (2022). Defects of the Innate Immune System and Related Immune Deficiencies.
6157 *Clin Rev Allergy Immunol* 63, 36-54.
- 6158 Alexovič, M., Sabo, J. and Longuespée, R. (2021). Automation of single-cell proteomic sample
6159 preparation. *PROTEOMICS* 21, 2100198.
- 6160 Aliverti, E., Lum, K., Johndrow, J.E. and Dunson, D.B. (2021). Removing the influence of group
6161 variables in high-dimensional predictive modelling. *J R Stat Soc Ser A Stat Soc* 184,
6162 791-811.
- 6163 Aliverti, E., Tilson, J.L., Filer, D.L., Babcock, B., Colaneri, A., Ocasio, J., Gershon, T.R.,
6164 Wilhelmsen, K.C. and Dunson, D.B. (2020). Projected t-SNE for batch correction.
6165 *Bioinformatics* 36, 3522-3527.
- 6166 Allen, C., Chang, Y., Ma, Q. and Chung, D. (2022). MAPLE: A Hybrid Framework for
6167 Multi-Sample Spatial Transcriptomics Data. *bioRxiv*, doi: 10.1101/2022.02.28.482296.
- 6168 Alon, S., Goodwin, D.R., Sinha, A., Wassie, A.T., Chen, F., Daugharty, E.R., Bando, Y., Kajita,
6169 A., Xue, A.G., Marrett, K., *et al.* (2021). Expansion sequencing: Spatially precise in situ
6170 transcriptomics in intact biological systems. *Science* 371, eaax2656.
- 6171 Alquicira-Hernandez, J., Sathe, A., Ji, H.P., Nguyen, Q. and Powell, J.E. (2019). scPred: accurate
6172 supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol*
6173 20, 264.
- 6174 Alvarez, M., Rahmani, E., Jew, B., Garske, K.M., Miao, Z., Benhammou, J.N., Ye, C.J., Pisegna,
6175 J.R., Pietilänen, K.H., Halperin, E., *et al.* (2020). Enhancing droplet-based single-nucleus
6176 RNA-seq resolution using the semi-supervised machine learning classifier DIEM. *Sci Rep* 10,
6177 11019.
- 6178 Amodio, M., van Dijk, D., Srinivasan, K., Chen, W.S., Mohsen, H., Moon, K.R., Campbell, A.,
6179 Zhao, Y., Wang, X., Venkataswamy, M., *et al.* (2019). Exploring single-cell data with deep
6180 multitasking neural networks. *Nat Methods* 16, 1139-1145.
- 6181 Anchang, B., Davis, K.L., Fienberg, H.G., Williamson, B.D., Bendall, S.C., Karacosta, L.G.,
6182 Tibshirani, R., Nolan, G.P. and Plevritis, S.K. (2018). DRUG-NEM: Optimizing drug
6183 combinations using single-cell perturbation response to account for intratumoral
6184 heterogeneity. *Proc Natl Acad Sci U S A* 115, E4294-e4303.
- 6185 Andersson, A., Andrusiová, Ž., Czarnewski, P., Li, X., Sundström, E. and Lundeberg, J. (2021a).

- 6186 A Landmark-based Common Coordinate Framework for Spatial Transcriptomics Data.
6187 bioRxiv, doi: 10.1101/2021.11.11.468178.
- 6188 Andersson, A., Bergenstrahle, J., Asp, M., Bergenstrahle, L., Jurek, A., Fernandez Navarro, J. and
6189 Lundeberg, J. (2020). Single-cell and spatial transcriptomics enables probabilistic inference
6190 of cell type topography. Commun Biol 3, 565.
- 6191 Andersson, A., Larsson, L., Stenbeck, L., Salmen, F., Ehinger, A., Wu, S.Z., Al-Eryani, G., Roden,
6192 D., Swarbrick, A., Borg, A., *et al.* (2021b). Spatial deconvolution of HER2-positive breast
6193 cancer delineates tumor-associated cell type interactions. Nat Commun 12, 6012.
- 6194 Andor, N., Lau, B.T., Catalanotti, C., Kumar, V., Sathe, A., Belhocine, K., Wheeler, T.D., Price,
6195 A.D., Song, M., Džakula, Ž., *et al.* (2020). Joint single cell DNA-Seq and RNA-Seq of
6196 cancer reveals subclonal signatures of genomic instability and gene expression. bioRxiv,
6197 doi:10.1101/445932.
- 6198 Andrews, T.S. and Hemberg, M. (2018). Identifying cell populations with scRNASeq. Mol
6199 Aspects Med 59, 114-122.
- 6200 Angelo, M., Bendall, S.C., Finck, R., Hale, M.B., Hitzman, C., Borowsky, A.D., Levenson, R.M.,
6201 Lowe, J.B., Liu, S.D., Zhao, S., *et al.* (2014). Multiplexed ion beam imaging of human breast
6202 tumors. Nat Med 20, 436-442.
- 6203 Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F.,
6204 Smallwood, S., Ponting, C.P., Voet, T., *et al.* (2016). Parallel single-cell sequencing links
6205 transcriptional and epigenetic heterogeneity. Nat Methods 13, 229-232.
- 6206 Angermueller, C., Lee, H.J., Reik, W. and Stegle, O. (2017). DeepCpG: accurate prediction of
6207 single-cell DNA methylation states using deep learning. Genome Biol 18, 67.
- 6208 Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J.,
6209 Abate, A.R., *et al.* (2019). Reference-based analysis of lung single-cell sequencing reveals a
6210 transitional profibrotic macrophage. Nat Immunol 20, 163-172.
- 6211 Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C. and Stegle, O.
6212 (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal
6213 single-cell data. Genome Biol 21, 111.
- 6214 Argelaguet, R., Clark, S.J., Mohammed, H., Stapel, L.C., Krueger, C., Kapourani, C.A.,
6215 Imaz-Rosshandler, I., Lohhoff, T., Xiang, Y., Hanna, C.W., *et al.* (2019). Multi-omics profiling
6216 of mouse gastrulation at single-cell resolution. Nature 576, 487-491.
- 6217 Argelaguet, R., Cuomo, A.S.E., Stegle, O. and Marioni, J.C. (2021). Computational principles and
6218 challenges in single-cell data integration. Nat Biotechnol 39, 1202-1215.
- 6219 Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W.
6220 and Stegle, O. (2018). Multi-Omics Factor Analysis-a framework for unsupervised
6221 integration of multi-omics data sets. Mol Syst Biol 14, e8124.
- 6222 Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X. and Garmire, L.X. (2019). DeepImpute: an
6223 accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data.
6224 Genome Biol 20, 211.
- 6225 Armingol, E., Baghdassarian, H.M., Martino, C., Perez-Lopez, A., Aamodt, C., Knight, R. and
6226 Lewis, N.E. (2022). Context-aware deconvolution of cell-cell communication with
6227 Tensor-cell2cell. Nat Commun 13, 3665.
- 6228 Arnol, D., Schapiro, D., Bodenmiller, B., Saez-Rodriguez, J. and Stegle, O. (2019). Modeling
6229 Cell-Cell Interactions from Spatial Molecular Data with Spatial Variance Component

- 6230 Analysis. Cell Rep 29, 202-211.e206.
- 6231 Arrastia, M.V., Jachowicz, J.W., Ollikainen, N., Curtis, M.S., Lai, C., Quinodoz, S.A., Selck, D.A.,
6232 Ismagilov, R.F. and Guttman, M. (2022). Single-cell measurement of higher-order 3D
6233 genome organization with scSPRITE. Nat Biotechnol 40, 64-73.
- 6234 Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski,
6235 K., Dwight, S.S., Eppig, J.T., *et al.* (2000). Gene ontology: tool for the unification of biology.
6236 The Gene Ontology Consortium. Nat Genet 25, 25-29.
- 6237 Ashuach, T., Gabitto, M.I., Jordan, M.I. and Yosef, N. (2021). MultiVI: deep generative model for
6238 the integration of multi-modal data. bioRxiv, doi: 10.1101/2021.08.20.457057.
- 6239 Ashuach, T., Reidenbach, D.A., Gayoso, A. and Yosef, N. (2022). PeakVI: A deep generative
6240 model for single-cell chromatin accessibility analysis. Cell Rep Methods 2, 100182.
- 6241 Asp, M., Giacomello, S., Larsson, L., Wu, C., Furth, D., Qian, X., Wardell, E., Custodio, J.,
6242 Reimegard, J., Salmen, F., *et al.* (2019). A Spatiotemporal Organ-Wide Gene Expression and
6243 Cell Atlas of the Developing Human Heart. Cell 179, 1647-1660.e1619.
- 6244 Assarsson, E., Lundberg, M., Holmquist, G., Björkesten, J., Thorsen, S.B., Ekman, D., Eriksson,
6245 A., Rennel Dickens, E., Ohlsson, S., Edfeldt, G., *et al.* (2014). Homogenous 96-plex PEA
6246 immunoassay exhibiting high sensitivity, specificity, and excellent scalability. PLoS One 9,
6247 e95192.
- 6248 Azodi, C.B., Zappia, L., Oshlack, A. and McCarthy, D.J. (2021). splatPop: simulating population
6249 scale single-cell RNA sequencing data. Genome Biol 22, 341.
- 6250 Bacher, R., Chu, L.F., Leng, N., Gasch, A.P., Thomson, J.A., Stewart, R.M., Newton, M. and
6251 Kendziorski, C. (2017). SCnorm: robust normalization of single-cell RNA-seq data. Nat
6252 Methods 14, 584-586.
- 6253 Bae, S., Choi, H. and Lee, D.S. (2021). Discovery of molecular features underlying the
6254 morphological landscape by integrating spatial transcriptomic data with deep features of
6255 tissue images. Nucleic Acids Res 49, e55.
- 6256 Bae, S., Na, K.J., Koh, J., Lee, D.S., Choi, H. and Kim, Y.T. (2022). CellDART: cell type
6257 inference by domain adaptation of single-cell and spatial transcriptomic data. Nucleic Acids
6258 Res 50, e57.
- 6259 Bai, L., Liang, J. and Cao, F. (2021). Semi-Supervised Clustering With Constraints of Different
6260 Types From Multiple Information Sources. IEEE Trans Pattern Anal Mach Intell 43,
6261 3247-3258.
- 6262 Bais, A.S. and Kostka, D. (2020). scds: computational annotation of doublets in single-cell RNA
6263 sequencing data. Bioinformatics 36, 1150-1158.
- 6264 Bandura, D.R., Baranov, V.I., Ornatsky, O.I., Antonov, A., Kinach, R., Lou, X., Pavlov, S.,
6265 Vorobiev, S., Dick, J.E. and Tanner, S.D. (2009). Mass cytometry: technique for real time
6266 single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass
6267 spectrometry. Anal Chem 81, 6813-6822.
- 6268 Bao, F., Deng, Y., Wan, S., Shen, S.Q., Wang, B., Dai, Q., Altschuler, S.J. and Wu, L.F. (2022).
6269 Integrative spatial analysis of cell morphologies and transcriptional states with MUSE. Nat
6270 Biotechnol 40, 1200-1209.
- 6271 Bar-Joseph, Z., Gitter, A. and Simon, I. (2012). Studying and modelling dynamic biological
6272 processes using time-series gene expression data. Nat Rev Genet 13, 552-564.
- 6273 Barkas, N., Petukhov, V., Nikolaeva, D., Lozinsky, Y., Demharter, S., Khodosevich, K. and

- 6274 Kharchenko, P.V. (2019). Joint analysis of heterogeneous single-cell RNA-seq dataset
 6275 collections. *Nat Methods* 16, 695-698.
- 6276 Bartosovic, M., Kabbe, M. and Castelo-Branco, G. (2021). Single-cell CUT&Tag profiles histone
 6277 modifications and transcription factors in complex tissues. *Nat Biotechnol* 39, 825-835.
- 6278 Barwinska, D., El-Achkar, T.M., Melo Ferreira, R., Syed, F., Cheng, Y.H., Winfree, S., Ferkowicz,
 6279 M.J., Hato, T., Collins, K.S., Dunn, K.W., *et al.* (2021). Molecular characterization of the
 6280 human kidney interstitium in health and disease. *Sci Adv* 7, eabd3359.
- 6281 Basile, G., Kahraman, S., Dirice, E., Pan, H., Dreyfuss, J.M. and Kulkarni, R.N. (2021). Using
 6282 single-nucleus RNA-sequencing to interrogate transcriptomic profiles of archived human
 6283 pancreatic islets. *Genome Med* 13, 128.
- 6284 Baslan, T., Kendall, J., Volyansky, K., McNamara, K., Cox, H., D'Italia, S., Ambrosio, F., Riggs,
 6285 M., Rodgers, L., Leotta, A., *et al.* (2020). Novel insights into breast cancer copy number
 6286 genetic heterogeneity revealed by single-cell genome sequencing. *Elife* 9, e51480.
- 6287 Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F. and Newell,
 6288 E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat
 6289 Biotechnol*, doi: 10.1038/nbt.4314.
- 6290 Bendall, S.C., Davis, K.L., Amir el, A.D., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld,
 6291 D.K., Nolan, G.P. and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression
 6292 and regulatory coordination in human B cell development. *Cell* 157, 714-725.
- 6293 Bendall, S.C., Simonds, E.F., Qiu, P., Amir el, A.D., Krutzik, P.O., Finck, R., Bruggner, R.V.,
 6294 Melamed, R., Trejo, A., Ornatsky, O.I., *et al.* (2011). Single-cell mass cytometry of
 6295 differential immune and drug responses across a human hematopoietic continuum. *Science*
 6296 332, 687-696.
- 6297 Bergen, V., Lange, M., Peidli, S., Wolf, F.A. and Theis, F.J. (2020). Generalizing RNA velocity to
 6298 transient cell states through dynamical modeling. *Nat Biotechnol* 38, 1408-1414.
- 6299 Bergen, V., Soldatov, R.A., Kharchenko, P.V. and Theis, F.J. (2021). RNA velocity-current
 6300 challenges and future perspectives. *Mol Syst Biol* 17, e10282.
- 6301 Bergenstrahle, L., He, B., Bergenstrahle, J., Abalo, X., Mirzazadeh, R., Thrane, K., Ji, A.L.,
 6302 Andersson, A., Larsson, L., Stakenborg, N., *et al.* (2022). Super-resolved spatial
 6303 transcriptomics by deep data fusion. *Nat Biotechnol* 40, 476-479.
- 6304 Bergman, H.M. and Lanekoff, I. (2017). Profiling and quantifying endogenous molecules in single
 6305 cells using nano-DESI MS. *Analyst* 142, 3639-3647.
- 6306 Bernstein, N.J., Fong, N.L., Lam, I., Roy, M.A., Hendrickson, D.G. and Kelley, D.R. (2020). Solo:
 6307 Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning. *Cell
 6308 Syst* 11, 95-101.e105.
- 6309 Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., Tokcan, N., Vanderburg,
 6310 C.R., Segerstolpe, A., Zhang, M., *et al.* (2021). Deep learning and alignment of spatially
 6311 resolved single-cell transcriptomes with Tangram. *Nat Methods* 18, 1352-1362.
- 6312 BinTayyash, N., Georgaka, S., John, S.T., Ahmed, S., Boukouvalas, A., Hensman, J. and Rattray,
 6313 M. (2021). Non-parametric modelling of temporal and spatial counts data from RNA-seq
 6314 experiments. *Bioinformatics* 37, 3788-3795.
- 6315 Blackburn, D.M., Lazure, F., Corchado, A.H., Perkins, T.J., Najafabadi, H.S. and Soleimani, V.D.
 6316 (2019). High-resolution genome-wide expression analysis of single myofibers using
 6317 SMART-Seq. *J Biol Chem* 294, 20097-20108.

- 6318 Blackburn, D.M., Lazure, F. and Soleimani, V.D. (2021). SMART approaches for genome-wide
 6319 analyses of skeletal muscle stem and niche cells. *Crit Rev Biochem Mol Biol* 56, 284-300.
- 6320 Blecher-Gonen, R., Barnett-Itzhaki, Z., Jaitin, D., Amann-Zalcenstein, D., Lara-Astiaso, D. and
 6321 Amit, I. (2013). High-throughput chromatin immunoprecipitation for genome-wide mapping
 6322 of in vivo protein-DNA interactions and epigenomic states. *Nat Protoc* 8, 539-554.
- 6323 Blei, D.M.L., J. D. A. (2007). A correlated topic model of science. *Ann. Appl Stat* 1, 17-35.
- 6324 Blondel, V.D., Guillaume, J.L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of
 6325 communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*
 6326 2008, P10008.
- 6327 Bloom, J.D. (2018). Estimating the frequency of multiplets in single-cell RNA sequencing from
 6328 cell-mixing experiments. *PeerJ* 6, e5578.
- 6329 Bock, C., Datlinger, P., Chardon, F. et al. (2022). High-content CRISPR screening. *Nat Rev
 6330 Methods Primers* 2, 8.
- 6331 Boileau, P., Hejazi, N.S. and Dudoit, S. (2020). Exploring high-dimensional biological data with
 6332 sparse contrastive principal component analysis. *Bioinformatics* 36, 3422-3430.
- 6333 Boisset, J.C., Vivié, J., Grün, D., Muraro, M.J., Lyubimova, A. and van Oudenaarden, A. (2018).
 6334 Mapping the physical network of cellular interactions. *Nat Methods* 15, 547-553.
- 6335 Bolger, A.M., Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina
 6336 sequence data. *Bioinformatics* 30, 2114-2120.
- 6337 Bolisetty, M.T., Stitzel, M.L. and Robson, P. (2017). CellView: Interactive exploration of high
 6338 dimensional single cell RNA-seq data. *bioRxiv*, doi:10.1101/123810.
- 6339 Borella, M., Martello, G., Risso, D. and Romualdi, C. (2021). PsiNorm: a scalable normalization
 6340 for single-cell RNA-seq data. *Bioinformatics* 38, 164-172.
- 6341 Boufea, K., Seth, S. and Batada, N.N. (2020). scID Uses Discriminant Analysis to Identify
 6342 Transcriptionally Equivalent Cell Types across Single-Cell RNA-Seq Data with Batch Effect.
 6343 *iScience* 23, 100914.
- 6344 Boyd, D.F., Allen, E.K., Randolph, A.G., Guo, X.J., Weng, Y., Sanders, C.J., Bajracharya, R., Lee,
 6345 N.K., Guy, C.S., Vogel, P., et al. (2020). Exuberant fibroblast activity compromises lung
 6346 function via ADAMTS4. *Nature* 587, 466-471.
- 6347 Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. and
 6348 Crawford, G.E. (2008). High-resolution mapping and characterization of open chromatin
 6349 across the genome. *Cell* 132, 311-322.
- 6350 Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V.,
 6351 Davie, K., Wouters, J. and Aerts, S. (2019). cisTopic: cis-regulatory topic modeling on
 6352 single-cell ATAC-seq data. *Nat Methods* 16, 397-400.
- 6353 Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L. (2016). Near-optimal probabilistic RNA-seq
 6354 quantification. *Nat Biotechnol* 34, 525-527.
- 6355 Breda, J., Zavolan, M. and van Nimwegen, E. (2021). Bayesian inference of gene expression
 6356 states from single-cell RNA-seq data. *Nat Biotechnol* 39, 1008-1016.
- 6357 Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B.,
 6358 Benes, V., Teichmann, S.A., Marioni, J.C., et al. (2013). Accounting for technical noise in
 6359 single-cell RNA-seq experiments. *Nat Methods* 10, 1093-1095.
- 6360 Brombacher, E., Hackenberg, M., Kreutz, C., Binder, H. and Treppner, M. (2022). The
 6361 performance of deep generative models for learning joint embeddings of single-cell

- 6362 multi-omics data. bioRxiv, doi: 10.1101/2022.06.06.494951.
- 6363 Browaeys, R., Saelens, W. and Saeys, Y. (2020). NicheNet: modeling intercellular communication
6364 by linking ligands to target genes. Nat Methods 17, 159-162.
- 6365 Brown, J., Ni, Z., Mohanty, C., Bacher, R. and Kendziorski, C. (2021). Normalization by
6366 distributional resampling of high throughput single-cell RNA-sequencing data.
6367 Bioinformatics 37, 4123-4128.
- 6368 Brüning, R.S., Tombor, L., Schulz, M.H., Dimmeler, S. and John, D. (2022). Comparative analysis
6369 of common alignment tools for single-cell RNA sequencing. Gigascience 11, giac001.
- 6370 Brunner, A.D., Thielert, M., Vasilopoulou, C., Ammar, C., Coscia, F., Mund, A., Hoerning, O.B.,
6371 Bache, N., Apalategui, A., Lubeck, M., *et al.* (2022). Ultra-high sensitivity mass spectrometry
6372 quantifies single-cell proteome changes upon perturbation. Mol Syst Biol 18, e10798.
- 6373 Budnik, B., Levy, E., Harmange, G. and Slavov, N. (2018). SCope-MS: mass spectrometry of
6374 single mammalian cells quantifies proteome heterogeneity during cell differentiation.
6375 Genome Biol 19, 161.
- 6376 Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. (2013). Transposition
6377 of native chromatin for fast and sensitive epigenomic profiling of open chromatin,
6378 DNA-binding proteins and nucleosome position. Nat Methods 10, 1213-1218.
- 6379 Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y.
6380 and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of
6381 regulatory variation. Nature 523, 486-490.
- 6382 Buhler, J. (2001). Efficient large-scale sequence comparison by locality-sensitive hashing.
6383 Bioinformatics 17, 419-428.
- 6384 Bullard, J.H., Purdom, E., Hansen, K.D. and Dudoit, S. (2010). Evaluation of statistical methods
6385 for normalization and differential expression in mRNA-Seq experiments. BMC
6386 Bioinformatics 11, 94.
- 6387 Burrows, M. and Wheeler, D.J. (1994). A Block-sorting Lossless Data Compression Algorithm.
6388 Src Research Report.
- 6389 C, P.E.d.S., Andronescu, M., Masud, T., Kabeer, F., Biele, J., Laks, E., Lai, D., Ye, P., Brimhall, J.,
6390 Wang, B., *et al.* (2020). Epiclomal: Probabilistic clustering of sparse single-cell DNA
6391 methylation data. PLoS Comput Biol 16, e1008270.
- 6392 Cabello-Aguilar, S., Alame, M., Kon-Sun-Tack, F., Fau, C., Lacroix, M. and Colinge, J. (2020).
6393 SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics.
6394 Nucleic Acids Res 48, e55.
- 6395 Cable, D.M., Murray, E., Zou, L.S., Goeva, A., Macosko, E.Z., Chen, F. and Irizarry, R.A. (2022).
6396 Robust decomposition of cell type mixtures in spatial transcriptomics. Nat Biotechnol 40,
6397 517-526.
- 6398 Cahill, J.F. and Kertesz, V. (2018). Automated Optically Guided System for Chemical Analysis of
6399 Single Plant and Algae Cells Using Laser Microdissection/Liquid Vortex Capture/Mass
6400 Spectrometry. Front Plant Sci 9, 1211.
- 6401 Cahill, J.F. and Kertesz, V. (2020). Laser Capture Microdissection-Liquid Vortex Capture Mass
6402 Spectrometry Metabolic Profiling of Single Onion Epidermis and Microalgae Cells. Methods
6403 Mol Biol 2064, 89-101.
- 6404 Cakir, B., Prete, M., Huang, N., van Dongen, S., Pir, P. and Kiselev, V.Y. (2020). Comparison of
6405 visualization tools for single-cell RNAseq data. NAR Genom Bioinform 2, lqaa052.

- 6406 Candelli, T., Schneider, P., Garrido Castro, P., Jones, L.A., Bodewes, E., Rockx-Brouwer, D.,
 6407 Pieters, R., Holstege, F.C.P., Margaritis, T. and Stam, R.W. (2022). Identification and
 6408 characterization of relapse-initiating cells in MLL-rearranged infant ALL by single-cell
 6409 transcriptomics. Leukemia 36, 58-67.
- 6410 Candès, E., Li, X., Ma, Y. and Wright, J.A. (2011). Robust principal component analysis? JACM
 6411 58, 11.
- 6412 Canete, N.P., Iyengar, S.S., Ormerod, J.T., Baharlou, H., Harman, A.N. and Patrick, E. (2022).
 6413 spicyR: spatial analysis of in situ cytometry data in R. Bioinformatics 38, 3099-3105.
- 6414 Cang, Z. and Nie, Q. (2020). Inferring spatial and signaling relationships between cells from
 6415 single cell transcriptomic data. Nat Commun 11, 2084.
- 6416 Cang, Z., Ning, X., Nie, A., Xu, M. and Zhang, J. (2021). SCAN-IT: Domain segmentation of
 6417 spatial transcriptomics images by graph neural network. BMVC 32, 406.
- 6418 Cannoodt, R., Saelens, W., Deconinck, L. and Saeys, Y. (2021). Spearheading future omics
 6419 analyses using dyngen, a multi-modal simulator of single cells. Nat Commun 12, 3942.
- 6420 Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M.,
 6421 McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., *et al.* (2018). Joint profiling of
 6422 chromatin accessibility and gene expression in thousands of single cells. Science 361,
 6423 1380-1385.
- 6424 Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan,
 6425 S.N., Steemers, F.J., *et al.* (2017). Comprehensive single-cell transcriptional profiling of a
 6426 multicellular organism. Science 357, 661-667.
- 6427 Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S.,
 6428 Christiansen, L., Steemers, F.J., *et al.* (2019). The single-cell transcriptional landscape of
 6429 mammalian organogenesis. Nature 566, 496-502.
- 6430 Cao, Y., Su, B., Guo, X., Sun, W., Deng, Y., Bao, L., Zhu, Q., Zhang, X., Zheng, Y., Geng, C., *et al.*
 6431 (2020a). Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-Throughput
 6432 Single-Cell Sequencing of Convalescent Patients' B Cells. Cell 182, 73-84.e16.
- 6433 Cao, Y., Wang, X. and Peng, G. (2020b). SCSA: A Cell Type Annotation Tool for Single-Cell
 6434 RNA-seq Data. Front Genet 11, 490.
- 6435 Cao, Z.-J. and Gao, G. (2022). Multi-omics single-cell data integration and regulatory inference
 6436 with graph-linked embedding. Nat Biotechnol 40, 1458-1466.
- 6437 Carter, B., Ku, W.L., Kang, J.Y., Hu, G., Perrie, J., Tang, Q. and Zhao, K. (2019). Mapping histone
 6438 modifications in low cell number and single cells using antibody-guided chromatin
 6439 fragmentation (ACT-seq). Nat Commun 10, 3747.
- 6440 Carter, B. and Zhao, K. (2021). The epigenetic basis of cellular heterogeneity. Nat Rev Genet 22,
 6441 235-250.
- 6442 Casey, M.J., Fliege, J., Sánchez-García, R.J. and MacArthur, B.D. (2023). An
 6443 information-theoretic approach to single cell sequencing analysis. bioRxiv, doi:
 6444 10.1101/2020.10.01.322255.
- 6445 Castro, D.C., Xie, Y.R., Rubakhin, S.S., Romanova, E.V. and Sweedler, J.V. (2021). Image-guided
 6446 MALDI mass spectrometry for high-throughput single-organelle characterization. Nat
 6447 Methods 18, 1233-1238.
- 6448 Chang, Y., He, F., Wang, J., Chen, S., Li, J., Liu, J., Yu, Y., Su, L., Ma, A., Allen, C., *et al.* (2022).
 6449 Define and visualize pathological architectures of human tissues from spatially resolved

- transcriptomics using deep learning. *Comput Struct Biotechnol J* 20, 4600-4617.
- Chatterton, Z., Lamichhane, P., Ahmadi Rastegar, D., Fitzpatrick, L., Lebhar, H., Marquis, C., Halliday, G. and Kwok, J.B. (2023). Single-cell DNA methylation sequencing by combinatorial indexing and enzymatic DNA methylation conversion. *Cell Biosci* 13, 2.
- Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., Qiu, X., Yang, J., Xu, J., Hao, S., *et al.* (2022). Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* 185, 1777-1792.e1721.
- Chen, C., Wu, C., Wu, L., Wang, X., Deng, M. and Xi, R. (2020a). scRMD: imputation for single cell RNA-seq data via robust matrix decomposition. *Bioinformatics* 36, 3156-3161.
- Chen, C., Xing, D., Tan, L., Li, H., Zhou, G., Huang, L. and Xie, X.S. (2017a). Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science* 356, 189-194.
- Chen, F., Lin, L., Zhang, J., He, Z., Uchiyama, K. and Lin, J.M. (2016a). Single-Cell Analysis Using Drop-on-Demand Inkjet Printing and Probe Electrospray Ionization Mass Spectrometry. *Anal Chem* 88, 4354-4360.
- Chen, G., Ning, B. and Shi, T. (2019a). Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front Genet* 10, 317.
- Chen, H., Albergante, L., Hsu, J.Y., Lareau, C.A., Lo Bosco, G., Guan, J., Zhou, S., Gorban, A.N., Bauer, D.E., Aryee, M.J., *et al.* (2019b). Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat Commun* 10, 1903.
- Chen, H., Lareau, C., Andreani, T., Vinyard, M.E., Garcia, S.P., Clement, K., Andrade-Navarro, M.A., Buenrostro, J.D. and Pinello, L. (2019c). Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol* 20, 241.
- Chen, H.I., Jin, Y., Huang, Y. and Chen, Y. (2016b). Detection of high variability in gene expression from single-cell RNA-seq profiling. *BMC Genomics* 17 Suppl 7, 508.
- Chen, J., Suo, S., Tam, P.P., Han, J.J., Peng, G. and Jing, N. (2017b). Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. *Nat Protoc* 12, 566-580.
- Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. and Zhuang, X. (2015a). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 348, aaa6090.
- Chen, L. and Zheng, S. (2018). BCseq: accurate single cell RNA-seq quantification with bias correction. *Nucleic Acids Res* 46, e82.
- Chen, M. and Zhou, X. (2018). VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biol* 19, 196.
- Chen, P.Y., Cokus, S.J. and Pellegrini, M. (2010). BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics* 11, 203.
- Chen, Q., Yan, G., Gao, M. and Zhang, X. (2015b). Ultrasensitive Proteome Profiling for 100 Living Cells by Direct Cell Injection, Online Digestion and Nano-LC-MS/MS Analysis. *Anal Chem* 87, 6674-6680.
- Chen, S., Lake, B.B. and Zhang, K. (2019d). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* 37, 1452-1457.
- Chen, S., Zhou, Y., Chen, Y. and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884-i890.
- Chen, W., Zhao, Y., Chen, X., Yang, Z., Xu, X., Bi, Y., Chen, V., Li, J., Choi, H., Ernest, B., *et al.* (2021a). A multicenter study benchmarking single-cell RNA sequencing technologies using

- 6494 reference samples. *Nat Biotechnol* 39, 1103-1114.
- 6495 Chen, W.T., Lu, A., Craessaerts, K., Pavie, B., Sala Frigerio, C., Corthout, N., Qian, X., Lalakova,
6496 J., Kuhnemund, M., Voytyuk, I., *et al.* (2020b). Spatial Transcriptomics and In Situ
6497 Sequencing to Study Alzheimer's Disease. *Cell* 182, 976-991.e919.
- 6498 Chen, Z., Yang, Z., Yuan, X., Zhang, X. and Hao, P. (2021b). scSensitiveGeneDefine: sensitive
6499 gene detection in single-cell RNA sequencing data by Shannon entropy. *BMC Bioinformatics*
6500 22, 211.
- 6501 Cheng, J., Zhang, J., Wu, Z. and Sun, X. (2021). Inferring microenvironmental regulation of gene
6502 expression from single-cell RNA sequencing data using scMLnet with an application to
6503 COVID-19. *Brief Bioinform* 22, 988-1005.
- 6504 Cheow, L.F., Courtois, E.T., Tan, Y., Viswanathan, R., Xing, Q., Tan, R.Z., Tan, D.S., Robson, P.,
6505 Loh, Y.H., Quake, S.R., *et al.* (2016). Single-cell multimodal profiling reveals cellular
6506 epigenetic heterogeneity. *Nat Methods* 13, 833-836.
- 6507 Cheung, T.K., Lee, C.-Y., Bayer, F.P., McCoy, A., Kuster, B. and Rose, C.M. (2021). Defining the
6508 carrier proteome limit for single-cell proteomics. *Nat Methods* 18, 76-83.
- 6509 Chidester, B., Zhou, T. and Ma, J. (2021). SPICEMIX: Integrative single-cell spatial modeling for
6510 inferring cell identity. *bioRxiv*, doi: 10.1101/2020.11.29.383067.
- 6511 Cho, C.S., Xi, J., Si, Y., Park, S.R., Hsu, J.E., Kim, M., Jun, G., Kang, H.M. and Lee, J.H. (2021).
6512 Microscopic examination of spatial transcriptome using Seq-Scope. *Cell* 184,
6513 3559-3572.e3522.
- 6514 Cho, H., Berger, B. and Peng, J. (2018a). Generalizable and Scalable Visualization of Single-Cell
6515 Data Using Neural Networks. *Cell Syst* 7, 185-191.e184.
- 6516 Cho, S.W., Xu, J., Sun, R., Mumbach, M.R., Carter, A.C., Chen, Y.G., Yost, K.E., Kim, J., He, J.,
6517 Nevins, S.A., *et al.* (2018b). Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA
6518 Boundary Element. *Cell* 173, 1398-1412.e1322.
- 6519 Chu, W.K., Edge, P., Lee, H.S., Bansal, V., Bafna, V., Huang, X. and Zhang, K. (2017).
6520 Ultraaccurate genome sequencing and haplotyping of single human cells. *Proc Natl Acad Sci
6521 U S A* 114, 12512-12517.
- 6522 Chua, R.L., Lukassen, S., Trump, S., Hennig, B.P., Wendisch, D., Pott, F., Debnath, O., Thürmann,
6523 L., Kurth, F., Völker, M.T., *et al.* (2020). COVID-19 severity correlates with airway
6524 epithelium-immune cell interactions identified by single-cell analysis. *Nat Biotechnol* 38,
6525 970-979.
- 6526 Chung, N.C. and Storey, J.D. (2015). Statistical significance of variables driving systematic
6527 variation in high-dimensional data. *Bioinformatics* 31, 545-554.
- 6528 Clark, S.J., Argelaguet, R., Kapourani, C.A., Stubbs, T.M., Lee, H.J., Alda-Catalinas, C., Krueger,
6529 F., Sanguinetti, G., Kelsey, G., Marioni, J.C., *et al.* (2018). scNMT-seq enables joint profiling
6530 of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun* 9,
6531 781.
- 6532 Clark, S.J., Smallwood, S.A., Lee, H.J., Krueger, F., Reik, W. and Kelsey, G. (2017).
6533 Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell
6534 bisulfite sequencing (scBS-seq). *Nat Protoc* 12, 534-547.
- 6535 Codeluppi, S., Borm, L.E., Zeisel, A., La Manno, G., van Lunteren, J.A., Svensson, C.I. and
6536 Linnarsson, S. (2018). Spatial organization of the somatosensory cortex revealed by
6537 osmFISH. *Nat Methods* 15, 932-935.

- 6538 Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson,
6539 S.F., Pellegrini, M. and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the
6540 Arabidopsis genome reveals DNA methylation patterning. *Nature* 452, 215-219.

6541 Collins, M., Dasgupta, S. and Schapire, R.E. (2002). A Generalization of Principal Component
6542 Analysis to the Exponential Family. In *Adv Neural Info Processing Sys* 14. MIT Press,
6543 Cambridge, MA.

6544 Conde, D., Triozzi, P.M., Balmant, K.M., Doty, A.L., Miranda, M., Boullosa, A., Schmidt, H.W.,
6545 Pereira, W.J., Dervinis, C. and Kirst, M. (2021). A robust method of nuclei isolation for
6546 single-cell RNA sequencing of solid tissues from the plant genus Populus. *PLoS One* 16,
6547 e0251149.

6548 Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W.,
6549 Marraffini, L.A., *et al.* (2013). Multiplex genome engineering using CRISPR/Cas systems.
6550 *Science* 339, 819-823.

6551 Cong, Y., Motamedchaboki, K., Misal, S.A., Liang, Y., Guise, A.J., Truong, T., Huguet, R., Plowey,
6552 E.D., Zhu, Y., Lopez-Ferrer, D., *et al.* (2020). Ultrasensitive single-cell proteomics workflow
6553 identifies >1000 protein groups per mammalian cell. *Chem Sci* 12, 1001-1006.

6554 Consortium, E.P., Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shores, N., Adrian, J.,
6555 Kawli, T., Davis, C.A., Dobin, A., *et al.* (2020a). Expanded encyclopaedias of DNA elements
6556 in the human and mouse genomes. *Nature* 583, 699-710.

6557 Consortium, E.P., Snyder, M.P., Gingeras, T.R., Moore, J.E., Weng, Z., Gerstein, M.B., Ren, B.,
6558 Hardison, R.C., Stamatoyannopoulos, J.A., Graveley, B.R., *et al.* (2020b). Perspectives on
6559 ENCODE. *Nature* 583, 693-698.

6560 Corces, M.R., Shcherbina, A., Kundu, S., Gloudemans, M.J., Fr  sard, L., Granja, J.M., Louie,
6561 B.H., Eulalio, T., Shams, S., Bagdatli, S.T., *et al.* (2020). Single-cell epigenomic analyses
6562 implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's
6563 diseases. *Nat Genet* 52, 1158-1168.

6564 Cortal, A., Martignetti, L., Six, E. and Rausell, A. (2021). Gene signature extraction and cell
6565 identity recognition at the single-cell level with Cell-ID. *Nat Biotechnol* 39, 1095-1102.

6566 Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V.,
6567 Gonzalez-Perez, A., Pearson, J., Sander, C., *et al.* (2015). Pathway and network analysis of
6568 cancer genomes. *Nat Methods* 12, 615-621.

6569 Crosse, E.I., Gordon-Keylock, S., Rybtsov, S., Binagui-Casas, A., Felchle, H., Nnadi, N.C.,
6570 Kirschner, K., Chandra, T., Tamagno, S., Webb, D.J., *et al.* (2020). Multi-layered Spatial
6571 Transcriptomics Identify Secretory Factors Promoting Human Hematopoietic Stem Cell
6572 Development. *Cell Stem Cell* 27, 822-839.e828.

6573 Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research.
6574 *Interjournal Complex Systems* 1695, 1-9.

6575 Ctortecka, C., Hartlmayr, D., Seth, A., Mendjan, S., Tourniaire, G. and Mechtler, K. (2022a). An
6576 automated workflow for multiplexed single-cell proteomics sample preparation at
6577 unprecedented sensitivity. *bioRxiv*, doi: 10.1101/2021.04.14.439828.

6578 Ctortecka, C., Kr  skov  , G., Stejskal, K., Penninger, J.M., Mendjan, S., Mechtler, K. and
6579 Stadlmann, J. (2022b). Comparative Proteome Signatures of Trace Samples by Multiplexed
6580 Data-Independent Acquisition. *Mol Cell Proteomics* 21, 100177.

6581 Cui, Y., Li, C., Jiang, Z., Zhang, S., Li, Q., Liu, X., Zhou, Y., Li, R., Wei, L., Li, L., *et al.* (2021).

- 6582 Single-cell transcriptome and genome analyses of pituitary neuroendocrine tumors. Neuro
 6583 Oncol 23, 1859-1871.
- 6584 Cui, Y., Zheng, Y., Liu, X., Yan, L., Fan, X., Yong, J., Hu, Y., Dong, J., Li, Q., Wu, X., *et al.*
 6585 (2019). Single-Cell Transcriptome Analysis Maps the Developmental Track of the Human
 6586 Heart. Cell Rep 26, 1934-1950.e1935.
- 6587 Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers,
 6588 F.J., Trapnell, C. and Shendure, J. (2015). Multiplex single cell profiling of chromatin
 6589 accessibility by combinatorial cellular indexing. Science 348, 910-914.
- 6590 Dago, A.E., Stepansky, A., Carlsson, A., Luttgen, M., Kendall, J., Baslan, T., Kolatkar, A., Wigler,
 6591 M., Bethel, K., Gross, M.E., *et al.* (2014). Rapid phenotypic and genomic change in response
 6592 to therapeutic pressure in prostate cancer inferred by high content analysis of single
 6593 circulating tumor cells. PLoS One 9, e101777.
- 6594 Daina, G., Ramos, L., Obradors, A., Rius, M., Martinez-Pasarell, O., Polo, A., Del Rey, J.,
 6595 Obradors, J., Benet, J. and Navarro, J. (2013). First successful double-factor PGD for Lynch
 6596 syndrome: monogenic analysis and comprehensive aneuploidy screening. Clin Genet 84,
 6597 70-73.
- 6598 Dal Molin, A. and Di Camillo, B. (2019). How to design a single-cell RNA-sequencing
 6599 experiment: pitfalls, challenges and perspectives. Brief Bioinform 20, 1384-1394.
- 6600 Danese, A., Richter, M.L., Chaichoompu, K., Fischer, D.S., Theis, F.J. and Colomé-Tatché, M.
 6601 (2021). EpiScanpy: integrated single-cell epigenomic analysis. Nat Commun 12, 5228.
- 6602 Dang, Y., Zhu, L., Yuan, P., Liu, Q., Guo, Q., Chen, X., Gao, S., Liu, X., Ji, S., Yuan, Y., *et al.*
 6603 (2023). Functional profiling of stage-specific proteome and translational transition across
 6604 human pre-implantation embryo development at a single-cell resolution. Cell Discovery 9,
 6605 10.
- 6606 Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster,
 6607 L.C., Kuchler, A., Alpar, D. and Bock, C. (2017). Pooled CRISPR screening with single-cell
 6608 transcriptome readout. Nat Methods 14, 297-301.
- 6609 Davies, D.L. and Bouldin, D.W. (1979). A cluster separation measure. IEEE Trans Pattern Anal
 6610 Mach Intell 1, 224-227.
- 6611 de Boer, C.G. and Regev, A. (2018). BROCKMAN: deciphering variance in epigenomic
 6612 regulators by k-mer factorization. BMC Bioinformatics 19, 253.
- 6613 de Kanter, J.K., Lijnzaad, P., Candelli, T., Margaritis, T. and Holstege, F.C.P. (2019). CHETAH: a
 6614 selective, hierarchical cell type identification method for single-cell RNA sequencing.
 6615 Nucleic Acids Res 47, e95.
- 6616 de Souza, N. (2012). The ENCODE project. Nat Methods 9, 1046.
- 6617 Dean, F.B., Nelson, J.R., Giesler, T.L. and Lasken, R.S. (2001). Rapid amplification of plasmid
 6618 and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle
 6619 amplification. Genome Res 11, 1095-1099.
- 6620 DeLaughter, D.M. (2018). The Use of the Fluidigm C1 for RNA Expression Analyses of Single
 6621 Cells. Curr Protoc Mol Biol 122, e55.
- 6622 Delcher, A.L., Phillippy, A., Carlton, J. and Salzberg, S.L. (2002). Fast algorithms for large-scale
 6623 genome alignment and comparison. Nucleic Acids Res 30, 2478-2483.
- 6624 Delmans, M. and Hemberg, M. (2016). Discrete distributional differential expression (D3E)--a
 6625 tool for gene expression analysis of single-cell RNA-seq data. BMC Bioinformatics 17, 110.

- 6626 Deng, Y., Bartosovic, M., Kukanja, P., Zhang, D., Liu, Y., Su, G., Enninful, A., Bai, Z.,
 6627 Castelo-Branco, G. and Fan, R. (2022a). Spatial-CUT&Tag: Spatially resolved chromatin
 6628 modification profiling at the cellular level. *Science* 375, 681-686.
- 6629 Deng, Y., Bartosovic, M., Ma, S., Zhang, D., Kukanja, P., Xiao, Y., Su, G., Liu, Y., Qin, X.,
 6630 Rosoklija, G.B., *et al.* (2022b). Spatial profiling of chromatin accessibility in mouse and
 6631 human tissues. *Nature* 609, 375-383.
- 6632 DePasquale, E.A.K., Schnell, D.J., Van Camp, P.J., Valiente-Aland í Í, Blaxall, B.C., Grimes, H.L.,
 6633 Singh, H. and Salomonis, N. (2019). DoubletDecon: Deconvoluting Doublets from
 6634 Single-Cell RNA-Sequencing Data. *Cell Rep* 29, 1718-1727.e1718.
- 6635 Dephoure, N. and Gygi, S.P. (2012). Hyperplexing: A Method for Higher-Order Multiplexed
 6636 Quantitative Proteomics Provides a Map of the Dynamic Response to Rapamycin in Yeast.
 6637 *Science Signaling* 5, rs2-rs2.
- 6638 Derkx, J., Leduc, A., Wallmann, G., Huffman, R.G., Willetts, M., Khan, S., Specht, H., Ralser, M.,
 6639 Demichev, V. and Slavov, N. (2022). Increasing the throughput of sensitive proteomics by
 6640 plexDIA. *bioRxiv*, doi: 10.1101/2021.11.03.467007.
- 6641 DeTomaso, D. and Yosef, N. (2016). FastProject: a tool for low-dimensional analysis of single-cell
 6642 RNA-Seq data. *BMC Bioinformatics* 17, 315.
- 6643 Dey, S.S., Kester, L., Spanjaard, B., Bienko, M. and van Oudenaarden, A. (2015). Integrated
 6644 genome and transcriptome sequencing of the same cell. *Nat Biotechnol* 33, 285-289.
- 6645 Dhainaut, M., Rose, S.A., Akturk, G., Wroblewska, A., Nielsen, S.R., Park, E.S., Buckup, M.,
 6646 Roudko, V., Pia, L., Sweeney, R., *et al.* (2022). Spatial CRISPR genomics identifies
 6647 regulators of the tumor microenvironment. *Cell* 185, 1223-1239.e1220.
- 6648 Dimitrov, D., Türei, D., Garrido-Rodriguez, M., Burmedi, P.L., Nagai, J.S., Boys, C., Ramirez
 6649 Flores, R.O., Kim, H., Szalai, B., Costa, I.G., *et al.* (2022). Comparison of methods and
 6650 resources for cell-cell communication inference from single-cell RNA-Seq data. *Nat
 6651 Commun* 13, 3224.
- 6652 Dimov, I.K., Kijanka, G., Park, Y., Ducr  e, J., Kang, T. and Lee, L.P. (2011). Integrated
 6653 microfluidic array plate (iMAP) for cellular and molecular analysis. *Lab Chip* 11, 2701-2710.
- 6654 Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., Wildberg, A. and Wang, W. (2015).
 6655 Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics* 31,
 6656 2225-2227.
- 6657 Ding, J., Shah, S. and Condon, A. (2016). densityCut: an efficient and versatile topological
 6658 approach for automatic clustering of biological data. *Bioinformatics* 32, 2567-2576.
- 6659 Ding, J., Sharon, N. and Bar-Joseph, Z. (2022). Temporal modelling using single-cell
 6660 transcriptomics. *Nat Rev Genet* 23, 355-368.
- 6661 Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D.,
 6662 Burks, T., Raychowdhury, R., *et al.* (2016). Perturb-Seq: Dissecting Molecular Circuits with
 6663 Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167, 1853-1866.e1817.
- 6664 Do, V.H. and Canzar, S. (2021). A generalization of t-SNE and UMAP to single-cell multimodal
 6665 omics. *Genome Biol* 22, 130.
- 6666 Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M.
 6667 and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29,
 6668 15-21.
- 6669 Dohmen, J., Baranovskii, A., Ronen, J., Uyar, B., Franke, V. and Akalin, A. (2021). Identifying

- 6670 tumor cells at the single cell level. bioRxiv, doi: 10.1101/2021.10.15.463909.
- 6671 Domínguez Conde, C., Xu, C., Jarvis, L.B., Rainbow, D.B., Wells, S.B., Gomes, T., Howlett, S.K.,
6672 Suchanek, O., Polanski, K., King, H.W., *et al.* (2022). Cross-tissue immune cell analysis
6673 reveals tissue-specific features in humans. *Science* 376, eabl5197.
- 6674 Dong, K. and Zhang, S. (2021). Joint reconstruction of cis-regulatory interaction networks across
6675 multiple tissues using single-cell chromatin accessibility data. *Brief Bioinform* 22, bbaa120.
- 6676 Dong, K. and Zhang, S. (2022). Deciphering spatial domains from spatially resolved
6677 transcriptomics with an adaptive graph attention auto-encoder. *Nat Commun* 13, 1739.
- 6678 Dong, R. and Yuan, G.C. (2020). GiniClust3: a fast and memory-efficient tool for rare cell type
6679 identification. *BMC Bioinformatics* 21, 158.
- 6680 Dong, R. and Yuan, G.C. (2021). SpatialDWLS: accurate deconvolution of spatial transcriptomic
6681 data. *Genome Biol* 22, 145.
- 6682 Dong, X., Tang, K., Xu, Y., Wei, H., Han, T. and Wang, C. (2022). Single-cell gene regulation
6683 network inference by large-scale data integration. *Nucleic Acids Res* 50, e126.
- 6684 Dries, R., Chen, J., Del Rossi, N., Khan, M.M., Sistig, A. and Yuan, G.C. (2021a). Advances in
6685 spatial transcriptomic data analysis. *Genome Res* 31, 1706-1718.
- 6686 Dries, R., Zhu, Q., Dong, R., Eng, C.L., Li, H., Liu, K., Fu, Y., Zhao, T., Sarkar, A., Bao, F., *et al.*
6687 (2021b). Giotto: a toolbox for integrative analysis and visualization of spatial expression data.
6688 *Genome Biol* 22, 78.
- 6689 Du, S., Zhai, L., Ye, S., Wang, L., Liu, M. and Tan, M. (2023). In-depth urinary and exosome
6690 proteome profiling analysis identifies novel biomarkers for diabetic kidney disease. *Sci China
6691 Life Sci* 66, 2587-2603.
- 6692 Duan, B., Zhou, C., Zhu, C., Yu, Y., Li, G., Zhang, S., Zhang, C., Ye, X., Ma, H., Qu, S., *et al.*
6693 (2019). Model-based understanding of single-cell CRISPR screening. *Nat Commun* 10, 2233.
- 6694 Duan, H., Li, F., Shang, J., Liu, J., Li, Y. and Liu, X. (2022). scVAEBGM: Clustering Analysis of
6695 Single-Cell ATAC-seq Data Using a Deep Generative Model. *Interdiscip Sci* 14, 917-928.
- 6696 Dueñas, M.E., Essner, J.J. and Lee, Y.J. (2017). 3D MALDI Mass Spectrometry Imaging of a
6697 Single Cell: Spatial Mapping of Lipids in the Embryonic Development of Zebrafish. *Sci Rep*
6698 7, 14946.
- 6699 Duncan, K.D., Bergman, H.M. and Lanekoff, I. (2017). A pneumatically assisted nanospray
6700 desorption electrospray ionization source for increased solvent versatility and enhanced
6701 metabolite detection from tissue. *Analyst* 142, 3424-3431.
- 6702 Duren, Z., Chang, F., Naqing, F., Xin, J., Liu, Q. and Wong, W.H. (2022). Regulatory analysis of
6703 single cell multiome gene expression and chromatin accessibility data with scREG. *Genome
6704 Biol* 23, 114.
- 6705 Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A.T., Chang, H.Y., Wang, Y. and
6706 Wong, W.H. (2018). Integrative analysis of single-cell genomics data by coupled nonnegative
6707 matrix factorizations. *Proc Natl Acad Sci U S A* 115, 7723-7728.
- 6708 Edsgard, D., Johnsson, P. and Sandberg, R. (2018). Identification of spatial expression trends in
6709 single-cell gene expression data. *Nat Methods* 15, 339-342.
- 6710 Efremova, M., Vento-Tormo, M., Teichmann, S.A. and Vento-Tormo, R. (2020). CellPhoneDB:
6711 inferring cell-cell communication from combined expression of multi-subunit ligand-receptor
6712 complexes. *Nat Protoc* 15, 1484-1506.
- 6713 Ellis, M.J., Ding, L., Shen, D., Luo, J., Suman, V.J., Wallis, J.W., Van Tine, B.A., Hoog, J.,

- 6714 Goiffon, R.J., Goldstein, T.C., *et al.* (2012). Whole-genome analysis informs breast cancer
6715 response to aromatase inhibition. *Nature* 486, 353-360.
- 6716 Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. and Heyn, H. (2021). SPOTlight: seeded NMF
6717 regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes.
6718 *Nucleic Acids Res* 49, e50.
- 6719 Elyanow, R., Dumitrascu, B., Engelhardt, B.E. and Raphael, B.J. (2020). netNMF-sc: leveraging
6720 gene-gene interactions for imputation and dimensionality reduction in single-cell expression
6721 analysis. *Genome Res* 30, 195-204.
- 6722 Emara, S., Amer, S., Ali, A., Abouleila, Y., Oga, A. and Masujima, T. (2017). Single-Cell
6723 Metabolomics. *Adv Exp Med Biol* 965, 323-343.
- 6724 Emmert-Buck, M.R., Bonner, R.F., Smith, P.D., Chuaqui, R.F., Zhuang, Z., Goldstein, S.R., Weiss,
6725 R.A. and Liotta, L.A. (1996). Laser capture microdissection. *Science* 274, 998-1001.
- 6726 Eng, C.L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C.,
6727 Yuan, G.C., *et al.* (2019). Transcriptome-scale super-resolved imaging in tissues by RNA
6728 seqFISH. *Nature* 568, 235-239.
- 6729 Engreitz, J.M., Haines, J.E., Perez, E.M., Munson, G., Chen, J., Kane, M., McDonel, P.E.,
6730 Guttman, M. and Lander, E.S. (2016). Local regulation of gene expression by lncRNA
6731 promoters, transcription and splicing. *Nature* 539, 452-455.
- 6732 Eraslan, G., Drokhlyansky, E., Anand, S., Fiskin, E., Subramanian, A., Slyper, M., Wang, J., Van
6733 Wittenberghe, N., Rouhana, J.M., Waldman, J., *et al.* (2022). Single-nucleus cross-tissue
6734 molecular reference maps toward understanding disease gene function. *Science* 376,
6735 eabl4290.
- 6736 Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S. and Theis, F.J. (2019). Single-cell RNA-seq
6737 denoising using a deep count autoencoder. *Nat Commun* 10, 390.
- 6738 Ernst, J. and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and
6739 characterization. *Nat Methods* 9, 215-216.
- 6740 Ester, M., Kriegel, H.P., Sander, J. and Xu, X.W. (1996). A Density-Based Algorithm for
6741 Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second*
6742 *International Conference on Knowledge Discovery and Data Mining*, 226-231.
- 6743 Evans, C., Hardin, J. and Stoebel, D.M. (2018). Selecting between-sample RNA-Seq
6744 normalization methods from the perspective of their assumptions. *Brief Bioinform* 19,
6745 776-792.
- 6746 Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal,
6747 B., Korninger, F., May, B., *et al.* (2018). The Reactome Pathway Knowledgebase. *Nucleic
6748 Acids Res* 46, D649-d655.
- 6749 Fan, H.C., Fu, G.K. and Fodor, S.P. (2015a). Expression profiling. Combinatorial labeling of
6750 single cells for gene expression cytometry. *Science* 347, 1258367.
- 6751 Fan, X., Zhang, X., Wu, X., Guo, H., Hu, Y., Tang, F. and Huang, Y. (2015b). Single-cell RNA-seq
6752 transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos.
6753 *Genome Biol* 16, 148.
- 6754 Fang, R., Preissl, S., Li, Y., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A.K., Zhou, X.,
6755 Xie, F., *et al.* (2021). Comprehensive analysis of single cell ATAC-seq data with SnapATAC.
6756 *Nat Commun* 12, 1337.
- 6757 Fang, X. and Ho, J.W.K. (2021). FlowGrid enables fast clustering of very large single-cell

- 6758 RNA-seq data. *Bioinformatics* 38, 282-283.
- 6759 Fanok, M.H., Sun, A., Fogli, L.K., Narendran, V., Eckstein, M., Kannan, K., Dolgalev, I., Lazaris,
6760 C., Heguy, A., Laird, M.E., *et al.* (2018). Role of Dysregulated Cytokine Signaling and
6761 Bacterial Triggers in the Pathogenesis of Cutaneous T-Cell Lymphoma. *J Invest Dermatol*
6762 138, 1116-1125.
- 6763 Farlik, M., Sheffield, N.C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J. and Bock, C.
6764 (2015). Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic
6765 cell-state dynamics. *Cell Rep* 10, 1386-1397.
- 6766 Fawkner-Corbett, D., Antanaviciute, A., Parikh, K., Jagielowicz, M., Geros, A.S., Gupta, T.,
6767 Ashley, N., Khamis, D., Fowler, D., Morrissey, E., *et al.* (2021). Spatiotemporal analysis of
6768 human intestinal development at single-cell resolution. *Cell* 184, 810-826.e823.
- 6769 Femino, A.M., Fay, F.S., Fogarty, K. and Singer, R.H. (1998). Visualization of single RNA
6770 transcripts in situ. *Science* 280, 585-590.
- 6771 Feng, D., Li, H., Xu, T., Zheng, F., Hu, C., Shi, X. and Xu, G. (2022). High-throughput single cell
6772 metabolomics and cellular heterogeneity exploration by inertial microfluidics coupled with
6773 pulsed electric field-induced electrospray ionization-high resolution mass spectrometry. *Anal
6774 Chim Acta* 1221, 340116.
- 6775 Feng, D., Whitehurst, C.E., Shan, D., Hill, J.D. and Yue, Y.G. (2019). Single Cell Explorer,
6776 collaboration-driven tools to leverage large-scale single cell RNA-seq data. *BMC Genomics*
6777 20, 676.
- 6778 Feng, Y. and Li, L.M. (2021). MUREN: a robust and multi-reference approach of RNA-seq
6779 transcript normalization. *BMC Bioinformatics* 22, 386.
- 6780 Ferguson, C.N., Fowler, J.W.M., Waxer, J.F., Gatti, R.A. and Loo, J.A. (2014). Mass
6781 Spectrometry-Based Tissue Imaging of Small Molecules. Springer International Publishing,
6782 doi: 10.1007/978-3-319-06068-2_12.
- 6783 Ferragina, P. and Manzini, G. (2001). An experimental study of an opportunistic index. In
6784 Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms, 269-278.
- 6785 Ferronika, P., van den Bos, H., Taudt, A., Spierings, D.C.J., Saber, A., Hiltermann, T.J.N., Kok, K.,
6786 Porubsky, D., van der Wekken, A.J., Timens, W., *et al.* (2017). Copy number alterations
6787 assessed at the single-cell level revealed mono- and polyclonal seeding patterns of distant
6788 metastasis in a small-cell lung cancer patient. *Ann Oncol* 28, 1668-1670.
- 6789 Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller,
6790 H.W., McElrath, M.J., Prlic, M., *et al.* (2015). MAST: a flexible statistical framework for
6791 assessing transcriptional changes and characterizing heterogeneity in single-cell RNA
6792 sequencing data. *Genome Biol* 16, 278.
- 6793 Fischer, D.S., Schaar, A.C. and Theis, F.J. (2021). Learning cell communication from spatial
6794 graphs of cells. *bioRxiv*, doi: 10.1101/2021.07.11.451750.
- 6795 Fischer, J. and Ayers, T. (2021). Single nucleus RNA-sequencing: how it's done, applications and
6796 limitations. *Emerg Top Life Sci* 5, 687-690.
- 6797 Fleck, J.S., Jansen, S.M.J., Wollny, D., Zenk, F., Seimiya, M., Jain, A., Okamoto, R., Santel, M.,
6798 He, Z., Camp, J.G., *et al.* (2023). Inferring and perturbing cell fate regulomes in human brain
6799 organoids. *Nature* 621, 365-372.
- 6800 Fleming, S.J., Marioni, J.C. and Babadi, M. (2019). CellBender remove-background: A deep
6801 generative model for unsupervised removal of background noise from scRNA-seq datasets.

- 6802 bioRxiv, doi: 10.1101/791699.
 6803 Flyamer, I.M., Gassler, J., Imakaev, M., Brandão, H.B., Ulianov, S.V., Abdennur, N., Razin, S.V.,
 6804 Mirny, L.A. and Tachibana-Konwalski, K. (2017). Single-nucleus Hi-C reveals unique
 6805 chromatin reorganization at oocyte-to-zygote transition. *Nature* 544, 110-114.
 6806 Fortunato, S. (2009). Community detection in graphs. *Phys Rep* 486, 75-174.
 6807 Francis, J.M., Zhang, C.Z., Maire, C.L., Jung, J., Manzo, V.E., Adalsteinsson, V.A., Homer, H.,
 6808 Haidar, S., Blumenstiel, B., Pedamallu, C.S., *et al.* (2014). EGFR variant heterogeneity in
 6809 glioblastoma resolved through single-nucleus sequencing. *Cancer Discov* 4, 956-971.
 6810 Franzén, O., Gan, L.M. and Björkegren, J.L.M. (2019). PanglaoDB: a web server for exploration
 6811 of mouse and human single-cell RNA sequencing data. *Database (Oxford)* 2019, baz046.
 6812 Frei, A.P., Bava, F.A., Zunder, E.R., Hsieh, E.W., Chen, S.Y., Nolan, G.P. and Gherardini, P.F.
 6813 (2016). Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat
 6814 Methods* 13, 269-275.
 6815 Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L. and
 6816 Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of
 6817 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* 89,
 6818 1827-1831.
 6819 Fu, H., Xu, H., Chong, K., Li, M., Ang, K.S., Lee, H.K., Ling, J., Chen, A., Shao, L., Liu, L., *et al.*
 6820 (2021a). Unsupervised Spatially Embedded Deep Representation of Spatial Transcriptomics.
 6821 bioRxiv, doi: 10.1101/2021.06.15.448542.
 6822 Fu, L., Zhang, L., Dollinger, E., Peng, Q., Nie, Q. and Xie, X. (2020). Predicting transcription
 6823 factor binding in single cells through deep learning. *Sci Adv* 6, eaba9031.
 6824 Fu, X., Sun, L., Chen, J.Y., Dong, R., Lin, Y., Palmiter, R.D., Lin, S. and Gu, L. (2021b).
 6825 Continuous Polony Gels for Tissue Mapping with High Resolution and RNA Capture
 6826 Efficiency. bioRxiv, doi: 10.1101/2021.03.17.435795.
 6827 Fu, Y., Li, C., Lu, S., Zhou, W., Tang, F., Xie, X.S. and Huang, Y. (2015). Uniform and accurate
 6828 single-cell sequencing based on emulsion whole-genome amplification. *Proc Natl Acad Sci U
 6829 S A* 112, 11923-11928.
 6830 Fujii, T., Matsuda, S., Tejedor, M.L., Esaki, T., Sakane, I., Mizuno, H., Tsuyama, N. and Masujima,
 6831 T. (2015). Direct metabolomics for plant cells by live single-cell mass spectrometry. *Nat
 6832 Protoc* 10, 1445-1456.
 6833 Fukunaga, K., and Hostetler, L. (1975). The estimation of the gradient of a density function, with
 6834 applications in pattern recognition. *IEEE Transactions on Information Theory* 21, 32-40.
 6835 Gabitto, M.I., Rasmussen, A., Wapinski, O., Allaway, K., Carriero, N., Fishell, G.J. and Bonneau,
 6836 R. (2020). Characterizing chromatin landscape from aggregate and single-cell genomic
 6837 assays using flexible duration modeling. *Nat Commun* 11, 747.
 6838 Galler, K., Bräutigam, K., Große, C., Popp, J. and Neugebauer, U. (2014). Making a big thing of a
 6839 small cell--recent advances in single cell analysis. *Analyst* 139, 1237-1273.
 6840 Gan, Y., Huang, X., Zou, G., Zhou, S. and Guan, J. (2022). Deep structural clustering for
 6841 single-cell RNA-seq data jointly through autoencoder and graph neural network. *Brief
 6842 Bioinform* 23, bbac018.
 6843 Gao, C., Liu, J., Kriebel, A.R., Preissl, S., Luo, C., Castanon, R., Sandoval, J., Rivkin, A., Nery,
 6844 J.R., Behrens, M.M., *et al.* (2021a). Iterative single-cell multi-omic integration using online
 6845 learning. *Nat Biotechnol* 39, 1000-1007.

- 6846 Gao, W., Ku, W.L., Pan, L., Perrie, J., Zhao, T., Hu, G., Wu, Y., Zhu, J., Ni, B. and Zhao, K.
 6847 (2021b). Multiplex indexing approach for the detection of DNase I hypersensitive sites in
 6848 single cells. Nucleic Acids Res 49, e56.
- 6849 Gao, X., Hu, D., Gogol, M. and Li, H. (2019). ClusterMap: compare multiple single cell RNA-Seq
 6850 datasets across different experimental conditions. Bioinformatics 35, 3038-3045.
- 6851 Garcia-Alonso, L., Handfield, L.F., Roberts, K., Nikolakopoulou, K., Fernando, R.C., Gardner, L.,
 6852 Woodhams, B., Arutyunyan, A., Polanski, K., Hoo, R., *et al.* (2021). Mapping the temporal
 6853 and spatial dynamics of the human endometrium in vivo and in vitro. Nat Genet 53,
 6854 1698-1711.
- 6855 Gauthier, M., Agniel, D., Thiébaut, R. and Hejblum, B.P. (2021). Distribution-free complex
 6856 hypothesis testing for single-cell RNA-seq differential expression analysis. bioRxiv, doi:
 6857 10.1101/2021.05.21.445165.
- 6858 Gawad, C., Koh, W. and Quake, S.R. (2014). Dissecting the clonal origins of childhood acute
 6859 lymphoblastic leukemia by single-cell genomics. Proc Natl Acad Sci U S A 111,
 6860 17947-17952.
- 6861 Gaydosik, A.M., Tabib, T., Domsic, R., Khanna, D., Lafyatis, R. and Fuschiotti, P. (2021).
 6862 Single-cell transcriptome analysis identifies skin-specific T-cell responses in systemic
 6863 sclerosis. Ann Rheum Dis 80, 1453-1460.
- 6864 Gebreyesus, S.T., Siyal, A.A., Kitata, R.B., Chen, E.S., Enkhbayar, B., Angata, T., Lin, K.I., Chen,
 6865 Y.J. and Tu, H.L. (2022). Streamlined single-cell proteomics by an integrated microfluidic
 6866 chip and data-independent acquisition mass spectrometry. Nat Commun 13, 37.
- 6867 Genshaft, A.S., Li, S., Gallant, C.J., Darmanis, S., Prakadan, S.M., Ziegler, C.G., Lundberg, M.,
 6868 Fredriksson, S., Hong, J., Regev, A., *et al.* (2016). Multiplexed, targeted profiling of
 6869 single-cell proteomes and transcriptomes in a single reaction. Genome Biol 17, 188.
- 6870 Gerlach, J.P., van Buggenum, J.A.G., Tanis, S.E.J., Hogeweg, M., Heuts, B.M.H., Muraro, M.J.,
 6871 Elze, L., Rivello, F., Rakszewska, A., van Oudenaarden, A., *et al.* (2019). Combined
 6872 quantification of intracellular (phospho-)proteins and transcriptomics from fixed single cells.
 6873 Sci Rep 9, 1469.
- 6874 Gerniers, A., Bricard, O. and Dupont, P. (2021). MicroCellClust: mining rare and highly specific
 6875 subpopulations from single-cell expression data. Bioinformatics 37, 3220-3227.
- 6876 Ghaddar, B. and De, S. (2022). Reconstructing physical cell interaction networks from single-cell
 6877 data using Neighbor-seq. Nucleic Acids Res 50, e82.
- 6878 Gierahn, T.M., Wadsworth, M.H., 2nd, Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune,
 6879 S., Love, J.C. and Shalek, A.K. (2017). Seq-Well: portable, low-cost RNA sequencing of
 6880 single cells at high throughput. Nat Methods 14, 395-398.
- 6881 Giesen, C., Wang, H.A., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schuffler, P.J.,
 6882 Grolimund, D., Buhmann, J.M., Brandt, S., *et al.* (2014). Highly multiplexed imaging of
 6883 tumor tissues with subcellular resolution by mass cytometry. Nat Methods 11, 417-422.
- 6884 Giladi, A., Cohen, M., Medaglia, C., Baran, Y., Li, B., Zada, M., Bost, P., Blecher-Gonen, R.,
 6885 Salame, T.M., Mayer, J.U., *et al.* (2020). Dissecting cellular crosstalk by sequencing
 6886 physically interacting cells. Nat Biotechnol 38, 629-637.
- 6887 Gogolewski, K., Sykulski, M., Chung, N.C. and Gambin, A. (2019). Truncated Robust Principal
 6888 Component Analysis and Noise Reduction for Single Cell RNA Sequencing Data. J Comput
 6889 Biol 26, 782-793.

- 6890 Goldstein, L.D., Chen, Y.J., Dunne, J., Mir, A., Hubschle, H., Guillory, J., Yuan, W., Zhang, J.,
 6891 Stinson, J., Jaiswal, B., *et al.* (2017). Massively parallel nanowell-based single-cell gene
 6892 expression profiling. *BMC Genomics* 18, 519.
- 6893 Gong, B., Zhou, Y. and Purdom, E. (2021). Cobolt: integrative analysis of multimodal single-cell
 6894 sequencing data. *Genome Biol* 22, 351.
- 6895 Gong, X., Zhao, Y., Cai, S., Fu, S., Yang, C., Zhang, S. and Zhang, X. (2014). Single cell analysis
 6896 with probe ESI-mass spectrometry: detection of metabolites at cellular and subcellular levels.
 6897 *Anal Chem* 86, 3809-3816.
- 6898 Gong, Y., Srinivasan, S.S., Zhang, R., Kessenbrock, K. and Zhang, J. (2022). scEpiLock: A
 6899 Weakly Supervised Learning Framework for cis-Regulatory Element Localization and
 6900 Variant Impact Quantification for Single-Cell Epigenetic Data. *Biomolecules* 12, 874.
- 6901 Gonzalez-Pena, V., Natarajan, S., Xia, Y., Klein, D., Carter, R., Pang, Y., Shaner, B., Annu, K.,
 6902 Putnam, D., Chen, W., *et al.* (2021). Accurate genomic variant detection in single cells with
 6903 primary template-directed amplification. *Proc Natl Acad Sci U S A* 118, e2024176118.
- 6904 Govek, K.W., Yamajala, V.S. and Camara, P.G. (2019). Clustering-independent analysis of
 6905 genomic data using spectral simplicial theory. *PLoS Comput Biol* 15, e1007509.
- 6906 Gralinska, E., Kohl, C., Sokhandan Fadakar, B. and Vingron, M. (2022). Visualizing
 6907 Cluster-specific Genes from Single-cell Transcriptomics Data Using Association Plots. *J Mol
 6908 Biol* 434, 167525.
- 6909 Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y. and Greenleaf,
 6910 W.J. (2021). ArchR is a scalable software package for integrative single-cell chromatin
 6911 accessibility analysis. *Nat Genet* 53, 403-411.
- 6912 Gravina, S., Ganapathi, S. and Vijg, J. (2015). Single-cell, locus-specific bisulfite sequencing
 6913 (SLBS) for direct detection of epimutations in DNA methylation patterns. *Nucleic Acids Res*
 6914 43, e93.
- 6915 Graving, J.M. and Couzin, I.D. (2020). VAE-SNE: a deep generative model for simultaneous
 6916 dimensionality reduction and clustering. *bioRxiv*, doi: 10.1101/2020.07.17.207993.
- 6917 Greguš, M., Kostas, J.C., Ray, S., Abbatiello, S.E. and Ivanov, A.R. (2020). Improved Sensitivity
 6918 of Ultralow Flow LC-MS-Based Proteomic Profiling of Limited Samples Using Monolithic
 6919 Capillary Columns and FAIMS Technology. *Anal Chem* 92, 14702-14712.
- 6920 Gross, A., Schoendube, J., Zimmermann, S., Steeb, M., Zengerle, R. and Koltay, P. (2015).
 6921 Technologies for Single-Cell Isolation. *Int J Mol Sci* 16, 16897-16919.
- 6922 Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., Dahmani, A.,
 6923 Lameiras, S., Reyal, F., Frenoy, O., *et al.* (2019). High-throughput single-cell ChIP-seq
 6924 identifies heterogeneity of chromatin states in breast cancer. *Nat Genet* 51, 1060-1066.
- 6925 Grothues, D., Cantor, C.R. and Smith, C.L. (1993). PCR amplification of megabase DNA with
 6926 tagged random primers (T-PCR). *Nucleic Acids Res* 21, 1321-1322.
- 6927 Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H. and van
 6928 Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell
 6929 types. *Nature* 525, 251-255.
- 6930 Grün, D. and van Oudenaarden, A. (2015). Design and Analysis of Single-Cell Sequencing
 6931 Experiments. *Cell* 163, 799-810.
- 6932 Grunau, C., Clark, S.J. and Rosenthal, A. (2001). Bisulfite genomic sequencing: systematic
 6933 investigation of critical experimental parameters. *Nucleic Acids Res* 29, E65-65.

- 6934 Gu, C. and Liu, Z. (2021). A network regularized linear model to infer spatial expression pattern
 6935 for single cells. bioRxiv, doi: 10.1101/2021.03.07.434296.
- 6936 Gu, H., Bock, C., Mikkelsen, T.S., Jäger, N., Smith, Z.D., Tomazou, E., Gnirke, A., Lander, E.S.
 6937 and Meissner, A. (2010). Genome-scale DNA methylation mapping of clinical samples at
 6938 single-nucleotide resolution. Nat Methods 7, 133-136.
- 6939 Gu, L., Li, X., Li, Z., Wang, Q., Zheng, K., Yu, G., Dai, C., Li, J., Zhao, B., Zhang, H., *et al.*
 6940 (2022a). Increasing the sensitivity, recovery, and integrality of spatially resolved proteomics
 6941 by LCM-MTA. bioRxiv, doi: 10.1101/2022.08.21.504675.
- 6942 Gu, L., Li, Z., Wang, Q., Zhang, H., Sun, Y., Li, C. and Wang, H. (2022b). An ultra-sensitive and
 6943 easy-to-use multiplexed single-cell proteomic analysis. bioRxiv, doi:
 6944 10.1101/2022.01.02.474723.
- 6945 Guo, H. and Li, J. (2021). scSorter: assigning cells to known cell types according to marker genes.
 6946 Genome Biol 22, 69.
- 6947 Guo, H., Zhu, P., Wu, X., Li, X., Wen, L. and Tang, F. (2013). Single-cell methylome landscapes
 6948 of mouse embryonic stem cells and early embryos analyzed using reduced representation
 6949 bisulfite sequencing. Genome Res 23, 2126-2135.
- 6950 Guo, M., Bao, E.L., Wagner, M., Whitsett, J.A. and Xu, Y. (2017). SLICE: determining cell
 6951 differentiation and lineage based on single cell entropy. Nucleic Acids Res 45, e54.
- 6952 Guo, M., Wang, H., Potter, S.S., Whitsett, J.A. and Xu, Y. (2015). SINCERA: A Pipeline for
 6953 Single-Cell RNA-Seq Profiling Analysis. PLoS Comput Biol 11, e1004575.
- 6954 Guo, X., Su, J., Zhou, H., Liu, C., Cao, J. and Li, L. (2019). Community Detection Based on
 6955 Genetic Algorithm Using Local Structural Similarity. IEEE Access 7, 134583-134600.
- 6956 Gupta, I., Collier, P.G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B.,
 6957 Smit, A.B., Sloan, S.A., *et al.* (2018). Single-cell isoform RNA sequencing characterizes
 6958 isoforms in thousands of cerebellar cells. Nat Biotechnol, doi: 10.1038/nbt.4259.
- 6959 Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S.R.,
 6960 Aguet, F., Gelfand, E., Ardlie, K., *et al.* (2017). Massively parallel single-nucleus RNA-seq
 6961 with DroNc-seq. Nat Methods 14, 955-958.
- 6962 Habib, N., Li, Y., Heidenreich, M., Swiech, L., Avraham-Davidi, I., Trombetta, J.J., Hession, C.,
 6963 Zhang, F. and Regev, A. (2016). Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare
 6964 adult newborn neurons. Science 353, 925-928.
- 6965 Hafemeister, C. and Satija, R. (2019). Normalization and variance stabilization of single-cell
 6966 RNA-seq data using regularized negative binomial regression. Genome Biol 20, 296.
- 6967 Hagié, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G., de Laat, W. and Forné, T.
 6968 (2007). Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nat
 6969 Protoc 2, 1722-1733.
- 6970 Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.J., Larsson, A.J.M.,
 6971 Faridani, O.R. and Sandberg, R. (2020). Single-cell RNA counting at allele and isoform
 6972 resolution using Smart-seq3. Nat Biotechnol 38, 708-714.
- 6973 Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F. and Theis, F.J. (2016). Diffusion pseudotime
 6974 robustly reconstructs lineage branching. Nat Methods 13, 845-848.
- 6975 Haghverdi, L., Lun, A.T.L., Morgan, M.D. and Marioni, J.C. (2018). Batch effects in single-cell
 6976 RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol
 6977 36, 421-427.

- 6978 Hahaut, V., Pavlinic, D., Carbone, W., Schuierer, S., Balmer, P., Quinodoz, M., Renner, M., Roma,
 6979 G., Cowan, C.S. and Picelli, S. (2022). Fast and highly sensitive full-length single-cell RNA
 6980 sequencing using FLASH-seq. *Nat Biotechnol* 40, 1447-1451.
- 6981 Haigis, K.M., Cichowski, K. and Elledge, S.J. (2019). Tissue-specificity in cancer: The rule, not
 6982 the exception. *Science* 363, 1150-1151.
- 6983 Han, K.Y., Kim, K.T., Joung, J.G., Son, D.S., Kim, Y.J., Jo, A., Jeon, H.J., Moon, H.S., Yoo, C.E.,
 6984 Chung, W., *et al.* (2018a). SIDR: simultaneous isolation and parallel sequencing of genomic
 6985 DNA and total RNA from single cells. *Genome Res* 28, 75-87.
- 6986 Han, L., Wu, H.J., Zhu, H., Kim, K.Y., Marjani, S.L., Riester, M., Euskirchen, G., Zi, X., Yang, J.,
 6987 Han, J., *et al.* (2017). Bisulfite-independent analysis of CpG island methylation enables
 6988 genome-scale stratification of single cells. *Nucleic Acids Res* 45, e77.
- 6989 Han, Q., Bradshaw, E.M., Nilsson, B., Hafler, D.A. and Love, J.C. (2010). Multidimensional
 6990 analysis of the frequencies and rates of cytokine secretion from single cells by quantitative
 6991 microengraving. *Lab on a Chip* 10, 1391.
- 6992 Han, W., Cheng, Y., Chen, J., Zhong, H., Hu, Z., Chen, S., Zong, L., Hong, L., Chan, T.F., King, I.,
 6993 *et al.* (2022). Self-supervised contrastive learning for integrative single cell RNA-seq data
 6994 analysis. *Brief Bioinform* 23, bbac377.
- 6995 Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., *et*
 6996 *al.* (2018b). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* 172, 1091-1107.e1017.
- 6997 Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., *et al.*
 6998 (2020). Construction of a human cell landscape at single-cell level. *Nature* 581, 303-309.
- 6999 Hao, L., Yu, S., Hao, H., Xin, H., Huan, T., Qiya, H., Longteng, W., Kang, X., Jingbo, G. and
 7000 Hebing, C. (2022). Inferring transcription factor regulatory networks from single-cell
 7001 ATAC-seq data based on graph neural networks. *Nature Machine Intelligence* 4, 389-400.
- 7002 Hao, M., Hua, K. and Zhang, X. (2021a). SOMDE: a scalable method for identifying spatially
 7003 variable genes with self-organizing map. *Bioinformatics* 37, 4392-4398.
- 7004 Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk,
 7005 A.J., Darby, C., Zager, M., *et al.* (2021b). Integrated analysis of multimodal single-cell data.
 7006 *Cell* 184, 3573-3587.e3529.
- 7007 Harada, A., Maehara, K., Handa, T., Arimura, Y., Nogami, J., Hayashi-Takanaka, Y., Shirahige, K.,
 7008 Kurumizaka, H., Kimura, H. and Ohkawa, Y. (2019). A chromatin integration labelling
 7009 method enables epigenomic profiling with lower input. *Nat Cell Biol* 21, 287-296.
- 7010 Harris, R.A., Wang, T., Coarfa, C., Nagarajan, R.P., Hong, C., Downey, S.L., Johnson, B.E., Fouse,
 7011 S.D., Delaney, A., Zhao, Y., *et al.* (2010). Comparison of sequencing-based methods to
 7012 profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat
 7013 Biotechnol* 28, 1097-1105.
- 7014 Hasanaj, E., Wang, J., Sarathi, A., Ding, J. and Bar-Joseph, Z. (2022). Interactive single-cell data
 7015 analysis using Cellar. *Nat Commun* 13, 1998.
- 7016 Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert,
 7017 D., Li, S., Livak, K.J., Rozenblatt-Rosen, O., *et al.* (2016). CEL-Seq2: sensitive
 7018 highly-multiplexed single-cell RNA-Seq. *Genome Biol* 17, 77.
- 7019 Hashimshony, T., Wagner, F., Sher, N. and Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by
 7020 multiplexed linear amplification. *Cell Rep* 2, 666-673.
- 7021 He, B., Bergenstrahle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, A., Maaskola, J.,

- 7022 Lundeberg, J. and Zou, J. (2020a). Integrating spatial gene expression and breast tumour
 7023 morphology via deep learning. *Nat Biomed Eng* 4, 827-834.
- 7024 He, D., Zakeri, M., Sarkar, H., Soneson, C., Srivastava, A. and Patro, R. (2022). Alevin-fry
 7025 unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data. *Nat
 7026 Methods* 19, 316-322.
- 7027 He, J., Zhao, F., Chen, B., Cui, N., Li, Z., Qin, J., Luo, L., Zhao, C. and Li, L. (2023). Alterations
 7028 in immune cell heterogeneities in the brain of aged zebrafish using single-cell resolution. *Sci
 7029 China Life Sci* 66, 1358-1378.
- 7030 He, S., Bhatt, R., Birditt, B., Brown, C., Brown, E., Chantranuvatana, K., Danaher, P., Dunaway,
 7031 D., Filanoski, B., Garrison, R.G., *et al.* (2021a). High-Plex Multiomic Analysis in FFPE
 7032 Tissue at Single-Cellular and Subcellular Resolution by Spatial Molecular Imaging. *bioRxiv*,
 7033 doi: 10.1101/2021.11.03.467020.
- 7034 He, X., Cai, D. and Niyogi, P. (2005). Laplacian Score for Feature Selection. In Proceedings of the
 7035 18th International Conference on Neural Information Processing Systems, 507-514.
- 7036 He, Y., Hariharan, M., Gorkin, D.U., Dickel, D.E., Luo, C., Castanon, R.G., Nery, J.R., Lee, A.Y.,
 7037 Zhao, Y., Huang, H., *et al.* (2020b). Spatiotemporal DNA methylome dynamics of the
 7038 developing mouse fetus. *Nature* 583, 752-759.
- 7039 He, Y., Tang, X., Huang, J., Ren, J., Zhou, H., Chen, K., Liu, A., Shi, H., Lin, Z., Li, Q., *et al.*
 7040 (2021b). ClusterMap for multi-scale clustering analysis of spatial gene expression. *Nat
 7041 Commun* 12, 5909.
- 7042 Hebenstreit, D. (2012). Methods, Challenges and Potentials of Single Cell RNA-seq. *Biology
 7043 (Basel)* 1, 658-667.
- 7044 Hedlund, E. and Deng, Q. (2018). Single-cell RNA sequencing: Technical advancements and
 7045 biological applications. *Mol Aspects Med* 59, 36-46.
- 7046 Herring, C.A., Banerjee, A., McKinley, E.T., Simmons, A.J., Ping, J., Roland, J.T., Franklin, J.L.,
 7047 Liu, Q., Gerdes, M.J., Coffey, R.J., *et al.* (2018). Unsupervised Trajectory Analysis of
 7048 Single-Cell RNA-Seq and Imaging Data Reveals Alternative Tuft Cell Origins in the Gut.
 7049 *Cell Syst* 6, 37-51.e39.
- 7050 Hicks, S.C., Liu, R., Ni, Y., Purdom, E. and Risso, D. (2021). mbkmeans: Fast clustering for single
 7051 cell data using mini-batch k-means. *PLoS Comput Biol* 17, e1008625.
- 7052 Hie, B., Bryson, B. and Berger, B. (2019). Efficient integration of heterogeneous single-cell
 7053 transcriptomes using Scanorama. *Nat Biotechnol* 37, 685-691.
- 7054 Hill, A.J., McFaline-Figueroa, J.L., Starita, L.M., Gasperini, M.J., Matreyek, K.A., Packer, J.,
 7055 Jackson, D., Shendure, J. and Trapnell, C. (2018). On the design of CRISPR-based single-cell
 7056 molecular screens. *Nat Methods* 15, 271-274.
- 7057 Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M.,
 7058 Cherniack, A.D., Thorsson, V., *et al.* (2018). Cell-of-Origin Patterns Dominate the Molecular
 7059 Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* 173, 291-304.e296.
- 7060 Hochgerner, H., Lönnberg, P., Hodge, R., Mikes, J., Heskol, A., Hubschle, H., Lin, P., Picelli, S.,
 7061 La Manno, G., Ratz, M., *et al.* (2017). STRT-seq-2i: dual-index 5' single cell and nucleus
 7062 RNA-seq on an addressable microwell array. *Sci Rep* 7, 16327.
- 7063 Hou, R., Denisenko, E. and Forrest, A.R.R. (2019). scMatch: a single-cell gene expression profile
 7064 annotation tool using reference datasets. *Bioinformatics* 35, 4688-4695.
- 7065 Hou, R., Denisenko, E., Ong, H.T., Ramilowski, J.A. and Forrest, A.R.R. (2020). Predicting

- 7066 cell-to-cell communication networks using NATMI. *Nat Commun* 11, 5011.
- 7067 Hou, Y., Fan, W., Yan, L., Li, R., Lian, Y., Huang, J., Li, J., Xu, L., Tang, F., Xie, X.S., *et al.*
7068 (2013). Genome analyses of single human oocytes. *Cell* 155, 1492-1506.
- 7069 Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y., *et al.*
7070 (2016). Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic
7071 heterogeneity in hepatocellular carcinomas. *Cell Res* 26, 304-319.
- 7072 Hu, H., Li, Z., Li, X., Yu, M. and Pan, X. (2022a). ScCAEs: deep clustering of single-cell
7073 RNA-seq via convolutional autoencoder embedding and soft K-means. *Brief Bioinform* 23,
7074 bbab321.
- 7075 Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D.J., Lee, E.B., Shinohara, R.T. and Li,
7076 M. (2021a). SpaGCN: Integrating gene expression, spatial location and histology to identify
7077 spatial domains and spatially variable genes by graph convolutional network. *Nat Methods*
7078 18, 1342-1351.
- 7079 Hu, J., Li, X., Hu, G., Lyu, Y., Susztak, K. and Li, M. (2020a). Iterative transfer learning with
7080 neural network for clustering and cell type classification in single-cell RNA-seq analysis. *Nat*
7081 *Mach Intell* 2, 607-618.
- 7082 Hu, J., Zhong, Y. and Shang, X. (2022b). A versatile and scalable single-cell data integration
7083 algorithm based on domain-adversarial and variational approximation. *Brief Bioinform* 23,
7084 bbab400.
- 7085 Hu, P., Zhang, W., Xin, H. and Deng, G. (2016a). Single Cell Isolation and Analysis. *Front Cell*
7086 *Dev Biol* 4, 116.
- 7087 Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., Zhu, X., Wang, C.Y., Xue, Z. and Fan, G.
7088 (2016b). Simultaneous profiling of transcriptome and DNA methylome from a single cell.
7089 *Genome Biol* 17, 88.
- 7090 Hu, Y., Peng, T., Gao, L. and Tan, K. (2021b). CytoTalk: De novo construction of signal
7091 transduction networks using single-cell transcriptomic data. *Sci Adv* 7, eabf1356.
- 7092 Hu, Z., Artibani, M., Alsaadi, A., Wietek, N., Morotti, M., Shi, T., Zhong, Z., Santana Gonzalez, L.,
7093 El-Sahhar, S., Carrami, E.M., *et al.* (2020b). The Repertoire of Serous Ovarian Cancer
7094 Non-genetic Heterogeneity Revealed by Single-Cell Sequencing of Normal Fallopian Tube
7095 Epithelial Cells. *Cancer Cell* 37, 226-242.e227.
- 7096 Hua, J., Liu, H., Zhang, B. and Jin, S. (2020). LAK: Lasso and K-means Based Single-cell
7097 RNA-seq Data Clustering Analysis. *IEEE Access* 8, 129679-129688.
- 7098 Huang, G., Li, G. and Cooks, R.G. (2011). Induced nanoelectrospray ionization for matrix-tolerant
7099 and high-throughput mass spectrometry. *Angew Chem Int Ed Engl* 50, 9907-9910.
- 7100 Huang, J., Yan, L., Fan, W., Zhao, N., Zhang, Y., Tang, F., Xie, X.S. and Qiao, J. (2014).
7101 Validation of multiple annealing and looping-based amplification cycle sequencing for
7102 24-chromosome aneuploidy screening of cleavage-stage embryos. *Fertil Steril* 102,
7103 1685-1691.
- 7104 Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J.I., Raj, A., Li, M.
7105 and Zhang, N.R. (2018a). SAVER: gene expression recovery for single-cell RNA sequencing.
7106 *Nat Methods* 15, 539-542.
- 7107 Huang, Q., Mao, S., Khan, M., Zhou, L. and Lin, J.M. (2018b). Dean flow assisted cell ordering
7108 system for lipid profiling in single-cells using mass spectrometry. *Chem Commun (Camb)* 54,
7109 2595-2598.

- 7110 Huang, Y. and Sanguinetti, G. (2021). BRIE2: computational identification of splicing phenotypes
7111 from single-cell transcriptomic experiments. *Genome Biol* 22, 251.
- 7112 Huang, Y. and Zhang, P. (2021). Evaluation of machine learning approaches for cell-type
7113 identification from single-cell transcriptomics data. *Brief Bioinform* 22, bbab035.
- 7114 Hughes, A.E., Magrini, V., Demeter, R., Miller, C.A., Fulton, R., Fulton, L.L., Eades, W.C., Elliott,
7115 K., Heath, S., Westervelt, P., *et al.* (2014a). Clonal architecture of secondary acute myeloid
7116 leukemia defined by single-cell sequencing. *PLoS Genet* 10, e1004462.
- 7117 Hughes, A.J., Spelke, D.P., Xu, Z., Kang, C.C., Schaffer, D.V. and Herr, A.E. (2014b). Single-cell
7118 western blotting. *Nat Methods* 11, 749-755.
- 7119 Hunter, M.V., Moncada, R., Weiss, J.M., Yanai, I. and White, R.M. (2021). Spatially resolved
7120 transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nat
7121 Commun* 12, 6278.
- 7122 Hunter, R.L., Actor, J.K., Hwang, S.A., Karev, V. and Jagannath, C. (2014). Pathogenesis of post
7123 primary tuberculosis: immunity and hypersensitivity in the development of cavities. *Ann Clin
7124 Lab Sci* 44, 365-387.
- 7125 Hutter, C. and Zenklusen, J.C. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond
7126 Its Data. *Cell* 173, 283-285.
- 7127 Hwang, B., Lee, J.H. and Bang, D. (2018). Single-cell RNA sequencing technologies and
7128 bioinformatics pipelines. *Exp Mol Med* 50, 1-14.
- 7129 Ianevski, A., Giri, A.K. and Aittokallio, T. (2022). Fully-automated and ultra-fast cell-type
7130 identification using specific marker combinations from single-cell transcriptomic data. *Nat
7131 Commun* 13, 1246.
- 7132 Ibáñez, A.J., Fagerer, S.R., Schmidt, A.M., Urban, P.L., Jefimovs, K., Geiger, P., Dechant, R.,
7133 Heinemann, M. and Zenobi, R. (2013). Mass spectrometry-based metabolomics of single
7134 yeast cells. *Proc Natl Acad Sci U S A* 110, 8790-8794.
- 7135 Ibáñez, A.J. and Svatos, A. (2020). Applications of MicroArrays for Mass Spectrometry (MAMS)
7136 in Single-Cell Metabolomics. *Methods Mol Biol* 2064, 73-88.
- 7137 Ilicic, T., Kim, J.K., Kolodziejczyk, A.A., Bagger, F.O., McCarthy, D.J., Marioni, J.C. and
7138 Teichmann, S.A. (2016). Classification of low quality cells from single-cell RNA-seq data.
7139 *Genome Biol* 17, 29.
- 7140 Isakova, A., Neff, N. and Quake, S.R. (2021). Single-cell quantification of a broad RNA spectrum
7141 reveals unique noncoding patterns associated with cell types and states. *Proc Natl Acad Sci U
7142 S A* 118, e2113568118.
- 7143 Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.B., Lönnerberg, P. and Linnarsson, S. (2011).
7144 Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq.
7145 *Genome Res* 21, 1160-1167.
- 7146 Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.B., Lönnerberg, P. and Linnarsson, S. (2012).
7147 Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nat Protoc* 7,
7148 813-828.
- 7149 Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P. and Linnarsson,
7150 S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods*
7151 11, 163-166.
- 7152 Jackson, C.A. and Vogel, C. (2022). New horizons in the stormy sea of multimodal single-cell data
7153 integration. *Mol Cell* 82, 248-259.

- 7154 Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A.,
 7155 Cohen, N., Jung, S., Tanay, A., *et al.* (2014). Massively parallel single-cell RNA-seq for
 7156 marker-free decomposition of tissues into cell types. *Science* 343, 776-779.
- 7157 Jaitin, D.A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T.M.,
 7158 Tanay, A., van Oudenaarden, A. and Amit, I. (2016). Dissecting Immune Circuits by Linking
 7159 CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* 167, 1883-1896.e1815.
- 7160 Jerby-Arnon, L., Shah, P., Cuoco, M.S., Rodman, C., Su, M.J., Melms, J.C., Leeson, R., Kanodia,
 7161 A., Mei, S., Lin, J.R., *et al.* (2018). A Cancer Cell Program Promotes T Cell Exclusion and
 7162 Resistance to Checkpoint Blockade. *Cell* 175, 984-997.e924.
- 7163 Ji, A.L., Rubin, A.J., Thrane, K., Jiang, S., Reynolds, D.L., Meyers, R.M., Guo, M.G., George,
 7164 B.M., Mollbrink, A., Bergenstrahle, J., *et al.* (2020a). Multimodal Analysis of Composition
 7165 and Spatial Architecture in Human Squamous Cell Carcinoma. *Cell* 182, 497-514.e422.
- 7166 Ji, Z. and Ji, H. (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell
 7167 RNA-seq analysis. *Nucleic Acids Res* 44, e117.
- 7168 Ji, Z., Zhou, W., Hou, W. and Ji, H. (2020b). Single-cell ATAC-seq signal extraction and
 7169 enhancement with SCATE. *Genome Biol* 21, 161.
- 7170 Ji, Z., Zhou, W. and Ji, H. (2017). Single-cell regulome data analysis by SCRAT. *Bioinformatics*
 7171 33, 2930-2932.
- 7172 Jia, C., Hu, Y., Kelly, D., Kim, J., Li, M. and Zhang, N.R. (2017). Accounting for technical noise
 7173 in differential expression analysis of single-cell RNA sequencing data. *Nucleic Acids Res* 45,
 7174 10978-10988.
- 7175 Jiang, H., Goulbourne, C.N., Tatar, A., Turlo, K., Wu, D., Beigneux, A.P., Grovenor, C.R., Fong,
 7176 L.G. and Young, S.G. (2014). High-resolution imaging of dietary lipids in cells and tissues by
 7177 NanoSIMS analysis. *J Lipid Res* 55, 2156-2166.
- 7178 Jiang, L., Chen, H., Pinello, L. and Yuan, G.C. (2016). GiniClust: detecting rare cell types from
 7179 single-cell gene expression data with Gini index. *Genome Biol* 17, 144.
- 7180 Jiang, R., Sun, T., Song, D. and Li, J.J. (2022a). Statistics or biology: the zero-inflation
 7181 controversy about scRNA-seq data. *Genome Biol* 23, 31
- 7182 Jiang Y, Gao X, Liu Y, Yan X, Shi H, Zhao R, Chen Z, Gao F, Zhao H, Zhao S (2024)
 7183 Cellular atlases of ovarian microenvironment alterations by diet and
 7184 genetically-induced obesity. *Sci. China Life Sci.*, 67, 51 - 6.liea
- 7185 Jiang, Y., Harigaya, Y., Zhang, Z., Zhang, H., Zang, C. and Zhang, N.R. (2022b). Nonparametric
 7186 single-cell multiomic characterization of trio relationships between transcription factors,
 7187 target genes, and cis-regulatory regions. *Cell Syst* 13, 737-751.e734.
- 7188 Jin, K., Li, B., Yan, H. and Zhang, X.F. (2022a). Imputing dropouts for single-cell RNA
 7189 sequencing based on multi-objective optimization. *Bioinformatics* 38, 3222-3230.
- 7190 Jin, S., Guerrero-Juarez, C.F., Zhang, L., Chang, I., Ramos, R., Kuan, C.H., Myung, P., Plikus,
 7191 M.V. and Nie, Q. (2021). Inference and analysis of cell-cell communication using CellChat.
 7192 *Nat Commun* 12, 1088.
- 7193 Jin, S., Zhang, L. and Nie, Q. (2020a). scAI: an unsupervised approach for the integrative analysis
 7194 of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol* 21, 25.
- 7195 Jin, W., Tang, Q., Wan, M., Cui, K., Zhang, Y., Ren, G., Ni, B., Sklar, J., Przytycka, T.M., Childs,
 7196 R., *et al.* (2015). Genome-wide detection of DNase I hypersensitive sites in single cells and
 7197 FFPE tissue samples. *Nature* 528, 142-146.

- 7198 Jin, X., Simmons, S.K., Guo, A., Shetty, A.S., Ko, M., Nguyen, L., Jokhi, V., Robinson, E., Oyler,
 7199 P., Curry, N., *et al.* (2020b). In vivo Perturb-Seq reveals neuronal and glial abnormalities
 7200 associated with autism risk genes. *Science* 370, eaaz6063.
- 7201 Jin, Z., Zhang, X., Dai, X., Huang, J., Hu, X., Zhang, J. and Shi, L. (2022b). InterCellDB: A
 7202 User-Defined Database for Inferring Intercellular Networks. *Adv Sci (Weinh)* 9, e2200045.
- 7203 Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007). Genome-wide mapping of in
 7204 vivo protein-DNA interactions. *Science* 316, 1497-1502.
- 7205 Jones, A., Townes, F.W., Li, D. and Engelhardt, B.E. (2022a). Alignment of spatial genomics and
 7206 histology data using deep Gaussian processes. *bioRxiv*, doi: 10.1101/2022.01.10.475692.
- 7207 Jones, R.C., Karkanias, J., Krasnow, M.A., Pisco, A.O., Quake, S.R., Salzman, J., Yosef, N.,
 7208 Bulthaup, B., Brown, P., Harper, W., *et al.* (2022b). The Tabula Sapiens: A multiple-organ,
 7209 single-cell transcriptomic atlas of humans. *Science* 376, eabl4896.
- 7210 Junker, J.P., Noël, E.S., Guryev, V., Peterson, K.A., Shah, G., Huisken, J., McMahon, A.P.,
 7211 Berezikov, E., Bakkers, J. and van Oudenaarden, A. (2014). Genome-wide RNA Tomography
 7212 in the zebrafish embryo. *Cell* 159, 662-675.
- 7213 Junntila, M.R. and de Sauvage, F.J. (2013). Influence of tumour micro-environment heterogeneity
 7214 on therapeutic response. *Nature* 501, 346-354.
- 7215 Kaminow, B., Yunusov, D. and Dobin, A. (2021). STARsolo: accurate, fast and versatile
 7216 mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv*, doi:
 7217 10.1101/2021.05.05.442755.
- 7218 Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E.,
 7219 Wong, S., Byrnes, L., Lanata, C.M., *et al.* (2018). Multiplexed droplet single-cell
 7220 RNA-sequencing using natural genetic variation. *Nat Biotechnol* 36, 89-94.
- 7221 Kang, J.B., Nathan, A., Weinand, K., Zhang, F., Millard, N., Rumker, L., Moody, D.B., Korsunsky,
 7222 I. and Raychaudhuri, S. (2021). Efficient and precise single-cell reference atlas mapping with
 7223 Symphony. *Nat Commun* 12, 5890.
- 7224 Kantlehner, M., Kirchner, R., Hartmann, P., Ellwart, J.W., Alunni-Fabbroni, M. and Schumacher,
 7225 A. (2011). A high-throughput DNA methylation analysis of a single cell. *Nucleic Acids Res*
 7226 39, e44.
- 7227 Kapourani, C.A., Argelaguet, R., Sanguinetti, G. and Vallejos, C.A. (2021). scMET: Bayesian
 7228 modeling of DNA methylation heterogeneity at single-cell resolution. *Genome Biol* 22, 114.
- 7229 Kapourani, C.A. and Sanguinetti, G. (2019). Melissa: Bayesian clustering and imputation of
 7230 single-cell methylomes. *Genome Biol* 20, 61.
- 7231 Karaïkos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., Kocks, C., Rajewsky, N.
 7232 and Zinzen, R.P. (2017). The *Drosophila* embryo at single-cell transcriptome resolution.
 7233 *Science* 358, 194-199.
- 7234 Karas, M., Bahr, U. and Dücks, T. (2000). Nano-electrospray ionization mass spectrometry:
 7235 addressing analytical problems beyond routine. *Fresenius J Anal Chem* 366, 669-676.
- 7236 Karayel, O., Virreira Winter, S., Padmanabhan, S., Kuras, Y.I., Vu, D.T., Tuncali, I., Merchant, K.,
 7237 Wills, A.M., Scherzer, C.R. and Mann, M. (2022). Proteome profiling of cerebrospinal fluid
 7238 reveals biomarker candidates for Parkinson's disease. *Cell Rep Med* 3, 100661.
- 7239 Kats, I., Vento-Tormo, R. and Stegle, O. (2021). SpatialDE2: Fast and localized variance
 7240 component analysis of spatial transcriptomics. *bioRxiv*, doi: 10.1101/2021.10.27.466045.
- 7241 Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K.

- 7242 and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and
 7243 single cells. *Nat Commun* 10, 1930.
- 7244 Kaymaz, Y., Ganglberger, F., Tang, M., Haslinger, C., Fernandez-Albert, F., Lawless, N. and
 7245 Sackton, T.B. (2021). HieRFIT: a hierarchical cell type classification tool for projections
 7246 from complex single-cell atlas datasets. *Bioinformatics* 37, 4431-4436.
- 7247 Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wahlby, C. and Nilsson, M. (2013).
 7248 In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Methods* 10, 857-860.
- 7249 Kebeschull, J.M., Richman, E.B., Ringach, N., Friedmann, D., Albaran, E., Kolluru, S.S., Jones,
 7250 R.C., Allen, W.E., Wang, Y., Cho, S.W., *et al.* (2020). Cerebellar nuclei evolved by repeatedly
 7251 duplicating a conserved cell-type set. *Science* 370, eabd5059.
- 7252 Kelly, R.T. (2020). Single-cell Proteomics: Progress and Prospects. *Mol Cell Proteomics* 19,
 7253 1739-1748.
- 7254 Kelly, T.K., Liu, Y., Lay, F.D., Liang, G., Berman, B.P. and Jones, P.A. (2012). Genome-wide
 7255 mapping of nucleosome positioning and DNA methylation within individual DNA molecules.
 7256 *Genome Res* 22, 2497-2506.
- 7257 Keren-Shaul, H., Kenigsberg, E., Jaitin, D.A., David, E., Paul, F., Tanay, A. and Amit, I. (2019).
 7258 MARS-seq2.0: an experimental and analytical pipeline for indexed sorting combined with
 7259 single-cell RNA sequencing. *Nat Protoc* 14, 1841-1862.
- 7260 Khan, S.A., Lehmann, R., Martinez-de-Morentin, X., Ruiz, A.M., Lagani, V., Kiani, N.A.,
 7261 Gomez-Cabrero, D. and Tegnér, J. (2022). scAEGAN: Unification of Single-Cell Genomics
 7262 Data by Adversarial Learning of Latent Space Correspondences. *bioRxiv*, doi:
 7263 10.1101/2022.04.19.488745.
- 7264 Kharchenko, P.V., Silberstein, L. and Scadden, D.T. (2014). Bayesian approach to single-cell
 7265 differential expression analysis. *Nat Methods* 11, 740-742.
- 7266 Khatri, P., Sirota, M. and Butte, A.J. (2012). Ten years of pathway analysis: current approaches
 7267 and outstanding challenges. *PLoS Comput Biol* 8, e1002375.
- 7268 Kim, C., Lee, H., Jeong, J., Jung, K. and Han, B. (2022). MarcoPolo: a method to discover
 7269 differentially expressed genes in single-cell RNA-seq data without depending on prior
 7270 clustering. *Nucleic Acids Res* 50, e71.
- 7271 Kim, D., Langmead, B. and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory
 7272 requirements. *Nat Methods* 12, 357-360.
- 7273 Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013). TopHat2:
 7274 accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.
 7275 *Genome Biol* 14, R36.
- 7276 Kim, H.J., Lin, Y., Geddes, T.A., Yang, J.Y.H. and Yang, P. (2020a). CiteFuse enables multi-modal
 7277 analysis of CITE-seq data. *Bioinformatics* 36, 4137-4143.
- 7278 Kim, T.H. and Ren, B. (2006). Genome-wide analysis of protein-DNA interactions. *Annu Rev
 7279 Genomics Hum Genet* 7, 81-102.
- 7280 Kim, T.H., Zhou, X. and Chen, M. (2020b). Demystifying "drop-outs" in single-cell UMI data.
 7281 *Genome Biol* 21, 196.
- 7282 Kind, J., Pagie, L., de Vries, S.S., Nahidazar, L., Dey, S.S., Bienko, M., Zhan, Y., Lajoie, B., de
 7283 Graaf, C.A., Amendola, M., *et al.* (2015). Genome-wide maps of nuclear lamina interactions
 7284 in single human cells. *Cell* 163, 134-147.
- 7285 Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N.,

- 7286 Reik, W., Barahona, M., Green, A.R., *et al.* (2017). SC3: consensus clustering of single-cell
 7287 RNA-seq data. *Nat Methods* 14, 483-486.
- 7288 Kiselev, V.Y., Yiu, A. and Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data
 7289 across data sets. *Nat Methods* 15, 359-362.
- 7290 Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A.
 7291 and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to
 7292 embryonic stem cells. *Cell* 161, 1187-1201.
- 7293 Klein, C.A., Schmidt-Kittler, O., Schardt, J.A., Pantel, K., Speicher, M.R. and Riethmüller, G.
 7294 (1999). Comparative genomic hybridization, loss of heterozygosity, and DNA sequence
 7295 analysis of single cells. *Proc Natl Acad Sci U S A* 96, 4494-4499.
- 7296 Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H.W., Li, T., Elmentait, R.,
 7297 Lomakin, A., Kedlian, V., Gayoso, A., *et al.* (2022). Cell2location maps fine-grained cell
 7298 types in spatial transcriptomics. *Nat Biotechnol* 40, 661-671.
- 7299 Knight, P., Gauthier, M.L., Pardo, C.E., Darst, R.P., Kapadia, K., Browder, H., Morton, E., Riva,
 7300 A., Kladde, M.P. and Bacher, R. (2021). Methylscaper: an R/Shiny app for joint visualization
 7301 of DNA methylation and nucleosome occupancy in single-molecule and single-cell data.
 7302 *Bioinformatics* 37, 4857-4859.
- 7303 Knouse, K.A., Wu, J., Whittaker, C.A. and Amon, A. (2014). Single cell sequencing reveals low
 7304 levels of aneuploidy across mammalian tissues. *Proc Natl Acad Sci U S A* 111, 13409-13414.
- 7305 Kobak, D. and Linderman, G.C. (2021). Initialization is critical for preserving global data
 7306 structure in both t-SNE and UMAP. *Nat Biotechnol* 39, 156-157.
- 7307 Kobayashi, H., Koike, T., Sakashita, A., Tanaka, K., Kumamoto, S. and Kono, T. (2016).
 7308 Repetitive DNA methylome analysis by small-scale and single-cell shotgun bisulfite
 7309 sequencing. *Genes Cells* 21, 1209-1222.
- 7310 Koenig, A.L., Shchukina, I., Amrute, J., Andhey, P.S., Zaitsev, K., Lai, L., Bajpai, G., Bredemeyer,
 7311 A., Smith, G., Jones, C., *et al.* (2022). Single-cell transcriptomics reveals cell-type-specific
 7312 diversification in human heart failure. *Nat Cardiovasc Res* 1, 263-280.
- 7313 Koike-Yusa, H., Li, Y., Tan, E.P., Velasco-Herrera Mdel, C. and Yusa, K. (2014). Genome-wide
 7314 recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library.
 7315 *Nat Biotechnol* 32, 267-273.
- 7316 Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C. and Teichmann, S.A. (2015). The
 7317 technology and biology of single-cell RNA sequencing. *Mol Cell* 58, 610-620.
- 7318 Kompauer, M., Heiles, S. and Spengler, B. (2017). Atmospheric pressure MALDI mass
 7319 spectrometry imaging of tissues and cells at 1.4-μm lateral resolution. *Nat Methods* 14,
 7320 90-96.
- 7321 Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner,
 7322 M., Loh, P.R. and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of
 7323 single-cell data with Harmony. *Nat Methods* 16, 1289-1296.
- 7324 Korthauer, K.D., Chu, L.F., Newton, M.A., Li, Y., Thomson, J., Stewart, R. and Kendziorski, C.
 7325 (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq
 7326 experiments. *Genome Biol* 17, 222.
- 7327 Kriebel, A.R. and Welch, J.D. (2022). UINMF performs mosaic integration of single-cell
 7328 multi-omic datasets using nonnegative matrix factorization. *Nat Commun* 13, 780.
- 7329 Krueger, F. and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for

- 7330 Bisulfite-Seq applications. Bioinformatics 27, 1571-1572.
- 7331 Ku, W.L., Nakamura, K., Gao, W., Cui, K., Hu, G., Tang, Q., Ni, B. and Zhao, K. (2019).
7332 Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone
7333 modification. Nat Methods 16, 323-325.
- 7334 Ku, W.L., Pan, L., Cao, Y., Gao, W. and Zhao, K. (2021). Profiling single-cell histone
7335 modifications using indexing chromatin immunocleavage sequencing. Genome Res 31,
7336 1831-1842.
- 7337 Kuchroo, M., Miyagishima, D.F., Steach, H.R., Godavarthi, A., Takeo, Y., Duy, P.Q., Barak, T.,
7338 Erson-Omay, E.Z., Youlten, S., Mishra-Gorur, K., *et al.* (2022). spARC recovers human
7339 glioma spatial signaling networks with graph filtering. bioRxiv, doi:
7340 10.1101/2022.08.24.505139.
- 7341 Kueckelhaus, J., von Ehr, J., Ravi, V.M., Will, P., Joseph, K., Beck, J., Hofmann, U.G., Delev, D.,
7342 Schnell, O. and Heiland, D.H. (2020). Inferring spatially transient gene expression pattern
7343 from spatial transcriptomic studies. bioRxiv, doi: 10.1101/2020.10.20.346544.
- 7344 Kumar, A., Ryan, A., Kitzman, J.O., Wemmer, N., Snyder, M.W., Sigurjonsson, S., Lee, C.,
7345 Banjevic, M., Zarutskie, P.W., Lewis, A.P., *et al.* (2015). Whole genome prediction for
7346 preimplantation genetic diagnosis. Genome Med 7, 35.
- 7347 Kumar, V.S. (2021). Seq-Well: Seeking A Simpler Way to Profile RNA from Single Cells. Clin
7348 Chem 67, 454-456.
- 7349 Kuppe, C., Ramirez Flores, R.O., Li, Z., Hayat, S., Levinson, R.T., Liao, X., Hannani, M.T.,
7350 Tanevski, J., Wunnemann, F., Nagai, J.S., *et al.* (2022). Spatial multi-omic map of human
7351 myocardial infarction. Nature 608, 766-777.
- 7352 Kurtenbach, S., Dollar, J.J., Cruz, A.M., Durante, M.A., Decatur, C.L. and Harbour, J.W. (2021).
7353 PieParty: visualizing cells from scRNA-seq data as pie charts. Life Sci Alliance 4,
7354 e202000986.
- 7355 La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K.,
7356 Kastriti, M.E., Lönnberg, P., Furlan, A., *et al.* (2018). RNA velocity of single cells. Nature
7357 560, 494-498.
- 7358 Labib, M. and Kelley, S.O. (2020). Single-cell analysis targeting the proteome. Nat Rev Chem 4,
7359 143-158.
- 7360 Lai, B., Gao, W., Cui, K., Xie, W., Tang, Q., Jin, W., Hu, G., Ni, B. and Zhao, K. (2018).
7361 Principles of nucleosome organization revealed by single-cell micrococcal nuclease
7362 sequencing. Nature 562, 281-285.
- 7363 Lake, B.B., Chen, S., Hoshi, M., Plongthongkum, N., Salamon, D., Knoten, A., Vijayan, A.,
7364 Venkatesh, R., Kim, E.H., Gao, D., *et al.* (2019). A single-nucleus RNA-sequencing pipeline
7365 to decipher the molecular anatomy and pathophysiology of human kidneys. Nat Commun 10,
7366 2832.
- 7367 Lake, B.B., Chen, S., Sos, B.C., Fan, J., Kaeser, G.E., Yung, Y.C., Duong, T.E., Gao, D., Chun, J.,
7368 Kharchenko, P.V., *et al.* (2018). Integrative single-cell analysis of transcriptional and
7369 epigenetic states in the human adult brain. Nat Biotechnol 36, 70-80.
- 7370 Laks, E., McPherson, A., Zahn, H., Lai, D., Steif, A., Brimhall, J., Biele, J., Wang, B., Masud, T.,
7371 Ting, J., *et al.* (2019). Clonal Decomposition and DNA Replication States Defined by Scaled
7372 Single-Cell Genome Sequencing. Cell 179, 1207-1221.e1222.
- 7373 Lal, A., Chiang, Z.D., Yakovenko, N., Duarte, F.M., Israeli, J. and Buenrostro, J.D. (2021). Deep

- 7374 learning-based enhancement of epigenomics data with AtacWorks. *Nat Commun* 12, 1507.
- 7375 Lall, S., Ray, S. and Bandyopadhyay, S. (2021). RgCop-A regularized copula based method for
7376 gene selection in single-cell RNA-seq data. *PLoS Comput Biol* 17, e1009464.
- 7377 Lan, F., Demaree, B., Ahmed, N. and Abate, A.R. (2017). Single-cell genome sequencing at
7378 ultra-high-throughput with microfluidic droplet barcoding. *Nat Biotechnol* 35, 640-646.
- 7379 Lance, C., Luecken, M.D., Burkhardt, D.B., Cannoodt, R., Rautenstrauch, P., Laddach, A.,
7380 Ubingazhibov, A., Cao, Z.-J., Deng, K., Khan, S., *et al.* (2022). Multimodal single cell data
7381 integration challenge: results and lessons learned. *bioRxiv*, doi: 10.1101/2022.04.11.487796.
- 7382 Langmead, B. and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*
7383 9, 357-359.
- 7384 Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009). Ultrafast and memory-efficient
7385 alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25.
- 7386 Langmead, B., Wilks, C., Antonescu, V. and Charles, R. (2019). Scaling read aligners to hundreds
7387 of threads on general-purpose processors. *Bioinformatics* 35, 421-432.
- 7388 Lareau, C.A., Duarte, F.M., Chew, J.G., Kartha, V.K., Burkett, Z.D., Kohlway, A.S., Pokholok, D.,
7389 Aryee, M.J., Steemers, F.J., Lebofsky, R., *et al.* (2019). Droplet-based combinatorial indexing
7390 for massive-scale single-cell chromatin accessibility. *Nat Biotechnol* 37, 916-924.
- 7391 Laurens, V.D.M. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine
7392 Learning Research* 9, 2579-2605.
- 7393 Lebrigand, K., Magnone, V., Barbry, P. and Waldmann, R. (2020). High throughput error corrected
7394 Nanopore single cell transcriptome sequencing. *Nat Commun* 11, 4025.
- 7395 Leduc, A., Huffman, R.G., Cantlon, J., Khan, S. and Slavov, N. (2022). Exploring functional
7396 protein covariation across single cells using nPOP. *Genome Biol* 23, 261.
- 7397 Lee, B., Namkoong, H., Yang, Y., Huang, H., Heller, D., Szot, G.L., Davis, M.M., Husain, S.Z.,
7398 Pandol, S.J., Bellin, M.D., *et al.* (2022). Single-cell sequencing unveils distinct immune
7399 microenvironments with CCR6-CCL20 crosstalk in human chronic pancreatitis. *Gut* 71,
7400 1831-1842.
- 7401 Lee, J.H., Daugharty, E.R., Scheiman, J., Kalhor, R., Ferrante, T.C., Terry, R., Turczyk, B.M.,
7402 Yang, J.L., Lee, H.S., Aach, J., *et al.* (2015). Fluorescent in situ sequencing (FISSEQ) of
7403 RNA for gene expression profiling in intact cells and tissues. *Nat Protoc* 10, 442-458.
- 7404 Legetth, O., Rodhe, J., Lang, S., Dhapola, P., Wallergård, M. and Soneji, S. (2021). CellexalVR: A
7405 virtual reality platform to visualize and analyze single-cell omics data. *iScience* 24, 103251.
- 7406 Lescroart, F., Wang, X., Lin, X., Swedlund, B., Gargouri, S., Sanchez-Danes, A., Moignard, V.,
7407 Dubois, C., Paulissen, C., Kinston, S., *et al.* (2018). Defining the earliest step of
7408 cardiovascular lineage segregation by single-cell RNA-seq. *Science* 359, 1177-1181.
- 7409 Leung, K., Zahn, H., Leaver, T., Konwar, K.M., Hanson, N.W., Pagé, A.P., Lo, C.C., Chain, P.S.,
7410 Hallam, S.J. and Hansen, C.L. (2012). A programmable droplet-based microfluidic device
7411 applied to multiparameter analysis of single microbes and microbial communities. *Proc Natl
7412 Acad Sci U S A* 109, 7665-7670.
- 7413 Leung, M.L., Davis, A., Gao, R., Casasent, A., Wang, Y., Sei, E., Vilar, E., Maru, D., Kopetz, S.
7414 and Navin, N.E. (2017). Single-cell DNA sequencing reveals a late-dissemination model in
7415 metastatic colorectal cancer. *Genome Res* 27, 1287-1299.
- 7416 Levy, E. and Slavov, N. (2018). Single cell protein analysis for systems biology. *Essays Biochem*
7417 62, 595-605.

- 7418 Li, B., Li, Y., Li, K., Zhu, L., Yu, Q., Cai, P., Fang, J., Zhang, W., Du, P., Jiang, C., *et al.* (2020a).
 7419 APEC: an accesson-based method for single-cell chromatin accessibility analysis. *Genome*
 7420 *Biol* 21, 116.
- 7421 Li, C., Fleck, J.S., Martins-Costa, C., Burkard, T.R., Thermann, J., Stuempflen, M., Peer, A.M.,
 7422 Vertesy, A., Littleboy, J.B., Esk, C., *et al.* (2023). Single-cell brain organoid screening
 7423 identifies developmental defects in autism. *Nature* 621, 373-380.
- 7424 Li, C., Liu, B., Kang, B., Liu, Z., Liu, Y., Chen, C., Ren, X. and Zhang, Z. (2020b). SciBet as a
 7425 portable and fast single cell type identifier. *Nat Commun* 11, 1818.
- 7426 Li, D., Velazquez, J.J., Ding, J., Hislop, J., Ebrahimkhani, M.R. and Bar-Joseph, Z. (2022a).
 7427 TraSig: inferring cell-cell interactions from pseudotime ordering of scRNA-Seq data.
 7428 *Genome Biol* 23, 73.
- 7429 Li, G., Fu, S., Wang, S., Zhu, C., Duan, B., Tang, C., Chen, X., Chuai, G., Wang, P. and Liu, Q.
 7430 (2022b). A deep generative model for multi-view profiling of single-cell RNA-seq and
 7431 ATAC-seq data. *Genome Biol* 23, 20.
- 7432 Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34,
 7433 3094-3100.
- 7434 Li, H., Brouwer, C.R. and Luo, W. (2022c). A universal deep neural network for in-depth cleaning
 7435 of single-cell RNA-Seq data. *Nat Commun* 13, 1901.
- 7436 Li, H., Courtois, E.T., Sengupta, D., Tan, Y., Chen, K.H., Goh, J.J.L., Kong, S.L., Chua, C., Hon,
 7437 L.K., Tan, W.S., *et al.* (2017). Reference component analysis of single-cell transcriptomes
 7438 elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet* 49, 708-718.
- 7439 Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler
 7440 transform. *Bioinformatics* 25, 1754-1760.
- 7441 Li, H., Ruan, J. and Durbin, R. (2008a). Mapping short DNA sequencing reads and calling
 7442 variants using mapping quality scores. *Genome Res* 18, 1851-1858.
- 7443 Li, H., Zhu, B., Xu, Z., Adams, T., Kaminski, N. and Zhao, H. (2021a). A Markov random field
 7444 model for network-based differential expression analysis of single-cell RNA-seq data. *BMC*
 7445 *Bioinformatics* 22, 524.
- 7446 Li, J., Sheng, Q., Shyr, Y. and Liu, Q. (2022d). scMRMA: single cell multiresolution marker-based
 7447 annotation. *Nucleic Acids Res* 50, e7.
- 7448 Li, K., Ouyang, Z., Chen, Y., Gagnon, J., Lin, D., Mingueneau, M., Chen, W., Sexton, D. and
 7449 Zhang, B. (2022e). Cellxgene VIP unleashes full power of interactive visualization and
 7450 integrative analysis of scRNA-seq, spatial transcriptomics, and multiome data. *bioRxiv*, doi:
 7451 10.1101/2020.08.28.270652.
- 7452 Li, L., Guo, F., Gao, Y., Ren, Y., Yuan, P., Yan, L., Li, R., Lian, Y., Li, J., Hu, B., *et al.* (2018a).
 7453 Single-cell multi-omics sequencing of human early embryos. *Nat Cell Biol* 20, 847-858.
- 7454 Li, L., Tang, H., Xia, R., Dai, H., Liu, R. and Chen, L. (2022f). Intrinsic entropy model for feature
 7455 selection of scRNA-seq data. *J Mol Cell Biol* 14, mjac008.
- 7456 Li, Q., Zhang, M., Xie, Y. and Xiao, G. (2021b). Bayesian Modeling of Spatial Molecular
 7457 Profiling Data via Gaussian Process. *Bioinformatics* 37, 4129-4136.
- 7458 Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008b). SOAP: short oligonucleotide alignment
 7459 program. *Bioinformatics* 24, 713-714.
- 7460 Li, R. and Yang, X. (2022). De novo reconstruction of cell interaction landscapes from single-cell
 7461 spatial transcriptome data with DeepLinc. *Genome Biol* 23, 124.

- 7462 Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K. and Wang, J. (2009). SOAP2: an
7463 improved ultrafast tool for short read alignment. Bioinformatics 25, 1966-1967.
- 7464 Li, S., Plouffe, B.D., Belov, A.M., Ray, S., Wang, X., Murthy, S.K., Karger, B.L. and Ivanov, A.R.
7465 (2015). An Integrated Platform for Isolation, Processing, and Mass Spectrometry-based
7466 Proteomic Profiling of Rare Cells in Whole Blood*. Molecular & Cellular Proteomics 14,
7467 1672-1683.
- 7468 Li, S., Su, K., Zhuang, Z., Qin, Q., Gao, L., Deng, Y., Liu, X., Hou, G., Wang, L., Hao, P., *et al.*
7469 (2022g). A simple, rapid, and practical method for single-cell proteomics based on
7470 mass-adaptive coating of synthetic peptides. Sci Bull (Beijing) 67, 581-584.
- 7471 Li, W., Xu, H., Xiao, T., Cong, L., Love, M.I., Zhang, F., Irizarry, R.A., Liu, J.S., Brown, M. and
7472 Liu, X.S. (2014). MAGeCK enables robust identification of essential genes from
7473 genome-scale CRISPR/Cas9 knockout screens. Genome Biol 15, 554.
- 7474 Li, W.V. and Li, J.J. (2018). An accurate and robust imputation method scImpute for single-cell
7475 RNA-seq data. Nat Commun 9, 997.
- 7476 Li, X., Lee, L., Abnousi, A., Yu, M., Liu, W., Huang, L., Li, Y. and Hu, M. (2022h). SnapHiC2: A
7477 computationally efficient loop caller for single cell Hi-C data. Comput Struct Biotechnol J 20,
7478 2778-2783.
- 7479 Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., Susztak, K., Reilly, M.P., Hu, G. and
7480 Li, M. (2020c). Deep learning enables accurate clustering with batch effect removal in
7481 single-cell RNA-seq analysis. Nat Commun 11, 2338.
- 7482 Li, X., Zeng, G., Li, A. and Zhang, Z. (2021c). DeTOKI identifies and characterizes the dynamics
7483 of chromatin TAD-like domains in a single cell. Genome Biol 22, 217.
- 7484 Li, Y., Ge, X., Peng, F., Li, W. and Li, J.J. (2022i). Exaggerated false positives by popular
7485 differential expression methods when analyzing human population samples. Genome Biol 23,
7486 79.
- 7487 Li, Y., Hu, X., Lin, R., Zhou, G., Zhao, L., Zhao, D., Zhang, Y., Li, W., Zhang, Y., Ma, P., *et al.*
7488 (2022j). Single-cell landscape reveals active cell subtypes and their interaction in the tumor
7489 microenvironment of gastric cancer. Theranostics 12, 3818-3833.
- 7490 Li, Y., Xu, X., Song, L., Hou, Y., Li, Z., Tsang, S., Li, F., Im, K.M., Wu, K., Wu, H., *et al.* (2012).
7491 Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in
7492 a muscle-invasive bladder cancer. Gigascience 1, 12.
- 7493 Li, Z.-Y., Huang, M., Wang, X.-K., Zhu, Y., Li, J.-S., Wong, C.C.L. and Fang, Q. (2018b).
7494 Nanoliter-Scale Oil-Air-Droplet Chip-Based Single Cell Proteomic Analysis. Anal Chem 90,
7495 5430-5438.
- 7496 Li, Z., Cheng, S., Lin, Q., Cao, W., Yang, J., Zhang, M., Shen, A., Zhang, W., Xia, Y., Ma, X., *et al.*
7497 (2021d). Single-cell lipidomics with high structural specificity by mass spectrometry. Nat
7498 Commun 12, 2869.
- 7499 Li, Z., Meisner, J. and Albrechtsen, A. (2022k). PCAOne: fast and accurate out-of-core PCA
7500 framework for large scale biobank data. bioRxiv, doi: 10.1101/2022.05.25.493261.
- 7501 Li, Z., Sun, C., Wang, F., Wang, X., Zhu, J., Luo, L., Ding, X., Zhang, Y., Ding, P., Wang, H., *et al.*
7502 (2022l). Molecular mechanisms governing circulating immune cell heterogeneity across
7503 different species revealed by single-cell sequencing. Clin Transl Med 12, e689.
- 7504 Li, Z. and Zhou, X. (2022). BASS: multi-scale and multi-sample analysis enables accurate cell
7505 type clustering and spatial domain detection in spatial transcriptomic studies. Genome Biol

- 7506 23, 168.
- 7507 Li, Z.Y., Huang, M., Wang, X.K., Zhu, Y., Li, J.S., Wong, C.C.L. and Fang, Q. (2018c).
7508 Nanoliter-Scale Oil-Air-Droplet Chip-Based Single Cell Proteomic Analysis. *Anal Chem* 90,
7509 5430-5438.
- 7510 Liang, L., Yu, J., Li, J., Li, N., Liu, J., Xiu, L., Zeng, J., Wang, T. and Wu, L. (2021a). Integration
7511 of scRNA-Seq and Bulk RNA-Seq to Analyse the Heterogeneity of Ovarian Cancer Immune
7512 Cells and Establish a Molecular Risk Model. *Front Oncol* 11, 711020.
- 7513 Liang, S., Mohanty, V., Dou, J., Miao, Q., Huang, Y., Müftüoğlu, M., Ding, L., Peng, W. and Chen,
7514 K. (2021b). Single-cell manifold-preserving feature selection for detecting rare cell
7515 populations. *Nat Comput Sci* 1, 374-384.
- 7516 Liang, Y., Acor, H., McCown, M.A., Nwosu, A.J., Boekweg, H., Axtell, N.B., Truong, T., Cong, Y.,
7517 Payne, S.H. and Kelly, R.T. (2021c). Fully Automated Sample Processing and Analysis
7518 Workflow for Low-Input Proteome Profiling. *Anal Chem* 93, 1658-1666.
- 7519 Liao, Y., Smyth, G.K. and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read
7520 mapping by seed-and-vote. *Nucleic Acids Res* 41, e108.
- 7521 Liao, Y., Smyth, G.K. and Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and
7522 better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res* 47, e47.
- 7523 Lin, C. and Bar-Joseph, Z. (2019). Continuous-state HMMs for modeling time-series single-cell
7524 RNA-Seq data. *Bioinformatics* 35, 4707-4715.
- 7525 Lin, G.L. and Hankenson, K.D. (2011). Integration of BMP, Wnt, and notch signaling pathways in
7526 osteoblast differentiation. *J Cell Biochem* 112, 3491-3501.
- 7527 Lin, H., Zhang, Z., Zhang, M.Q., Ma, B. and Li, M. (2008). ZOOM! Zillions of oligos mapped.
7528 *Bioinformatics* 24, 2431-2437.
- 7529 Lin, L. and Zhang, L. (2022). Joint analysis of scATAC-seq datasets using epiConv. *BMC
7530 Bioinformatics* 23, 309.
- 7531 Lin, Y., Wu, T.Y., Wan, S., Yang, J.Y.H., Wong, W.H. and Wang, Y.X.R. (2022). scJoint integrates
7532 atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nat Biotechnol* 40,
7533 703-710.
- 7534 Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S. and Kluger, Y. (2019). Fast
7535 interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat
7536 Methods* 16, 243-245.
- 7537 Linderman, G.C., Zhao, J., Roulis, M., Bielecki, P., Flavell, R.A., Nadler, B. and Kluger, Y. (2022).
7538 Zero-preserving imputation of single-cell RNA-seq data. *Nat Commun* 13, 192.
- 7539 Liscovitch-Brauer, N., Montalbano, A., Deng, J., Méndez-Mancilla, A., Wessels, H.H., Moss, N.G.,
7540 Kung, C.Y., Sookdeo, A., Guo, X., Geller, E., *et al.* (2021). Profiling the genetic determinants
7541 of chromatin accessibility with scalable single-cell CRISPR screens. *Nat Biotechnol* 39,
7542 1270-1277.
- 7543 Littman, R., Hemminger, Z., Foreman, R., Arneson, D., Zhang, G., Gomez-Pinilla, F., Yang, X.
7544 and Wollman, R. (2021). Joint cell segmentation and cell type annotation for spatial
7545 transcriptomics. *Mol Syst Biol* 17, e10108.
- 7546 Liu, H., Zhou, J., Tian, W., Luo, C., Bartlett, A., Aldridge, A., Lucero, J., Osteen, J.K., Nery, J.R.,
7547 Chen, H., *et al.* (2021). DNA methylation atlas of the mouse brain at single-cell resolution.
7548 *Nature* 598, 120-128.
- 7549 Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J.,

- 7550 Benz, C.C., Levine, D.A., Lee, A.V., *et al.* (2018a). An Integrated TCGA Pan-Cancer Clinical
 7551 Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* 173, 400-416.e411.
- 7552 Liu, L., Chen, D., Wang, J. and Chen, J. (2020a). Advances of Single-Cell Protein Analysis. *Cells*
 7553 9, 1271.
- 7554 Liu, M., Liu, Y., Di, J., Su, Z., Yang, H., Jiang, B., Wang, Z., Zhuang, M., Bai, F. and Su, X.
 7555 (2017). Multi-region and single-cell sequencing reveal variable genomic heterogeneity in
 7556 rectal cancer. *BMC Cancer* 17, 787.
- 7557 Liu, N., Liu, L. and Pan, X. (2014). Single-cell analysis of the transcriptome and its application in
 7558 the characterization of stem cells and early embryos. *Cell Mol Life Sci* 71, 2707-2715.
- 7559 Liu, R., Pan, N., Zhu, Y. and Yang, Z. (2018b). T-Probe: An Integrated Microscale Device for
 7560 Online In Situ Single Cell Analysis and Metabolic Profiling Using Mass Spectrometry. *Anal*
 7561 *Chem* 90, 11078-11085.
- 7562 Liu, R. and Yang, Z. (2021). Single cell metabolomics using mass spectrometry: Techniques and
 7563 data analysis. *Anal Chim Acta* 1143, 124-134.
- 7564 Liu, T. and Wang, Z. (2022). scHiCEmbed: Bin-Specific Embeddings of Single-Cell Hi-C Data
 7565 Using Graph Auto-Encoders. *Genes (Basel)* 13, 1048.
- 7566 Liu, W., Liao, X., Luo, Z., Yang, Y., Lau, M.C., Jiao, Y., Shi, X., Zhai, W., Ji, H., Yeong, J., *et al.*
 7567 (2022). Probabilistic embedding, clustering, and alignment for integrating spatial
 7568 transcriptomics data with PRECAST. *bioRxiv*, doi: 10.1101/2022.06.26.497672.
- 7569 Liu, Y., Chen, X., Zhang, Y. and Liu, J. (2019). Advancing single-cell proteomics and
 7570 metabolomics with microfluidic technologies. *Analyst* 144, 846-858.
- 7571 Liu, Y., Li, H.D., Xu, Y., Liu, Y.W., Peng, X. and Wang, J. (2023). IsoCell: An Approach to
 7572 Enhance Single Cell Clustering by Integrating Isoform-Level Expression Through
 7573 Orthogonal Projection. *IEEE/ACM Trans Comput Biol Bioinform* 20, 465-475.
- 7574 Liu, Y., Yang, M., Deng, Y., Su, G., Enninfu, A., Guo, C.C., Tebaldi, T., Zhang, D., Kim, D., Bai,
 7575 Z., *et al.* (2020b). High-Spatial-Resolution Multi-Omics Sequencing via Deterministic
 7576 Barcoding in Tissue. *Cell* 183, 1665-1681.e1618.
- 7577 Lodato, M.A., Woodworth, M.B., Lee, S., Evrony, G.D., Mehta, B.K., Karger, A., Lee, S.,
 7578 Chittenden, T.W., D'Gama, A.M., Cai, X., *et al.* (2015). Somatic mutation in single human
 7579 neurons tracks developmental and transcriptional history. *Science* 350, 94-98.
- 7580 Lohoff, T., Ghazanfar, S., Missarova, A., Koulena, N., Pierson, N., Griffiths, J.A., Bardot, E.S.,
 7581 Eng, C.L., Tyser, R.C.V., Argelaguet, R., *et al.* (2022). Integration of spatial and single-cell
 7582 transcriptomic data elucidates mouse organogenesis. *Nat Biotechnol* 40, 74-85.
- 7583 Lombard-Banek, C., Moody, S.A., Manzini, M.C. and Nemes, P. (2019). Microsampling Capillary
 7584 Electrophoresis Mass Spectrometry Enables Single-Cell Proteomics in Complex Tissues:
 7585 Developing Cell Clones in Live *Xenopus laevis* and Zebrafish Embryos. *Anal Chem* 91,
 7586 4797-4805.
- 7587 Lombard-Banek, C., Moody, S.A. and Nemes, P. (2016). Single-Cell Mass Spectrometry for
 7588 Discovery Proteomics: Quantifying Translational Cell Heterogeneity in the 16-Cell Frog
 7589 (*Xenopus*) Embryo. *Angew Chem Int Ed Engl* 55, 2454-2458.
- 7590 Long, Y., Ang, K.S., Li, M., Chong, K.L.K., Sethi, R., Zhong, C., Xu, H., Ong, Z., Sachaphibulkij,
 7591 K., Chen, A., *et al.* (2023). Spatially informed clustering, integration, and deconvolution of
 7592 spatial transcriptomics with GraphST. *Nat Commun* 14, 1155.
- 7593 Lopez-Delisle, L. and Delisle, J.B. (2022). baredSC: Bayesian approach to retrieve expression

- 7594 distribution of single-cell data. BMC Bioinformatics 23, 36.
- 7595 Lopez, R., Li, B., Keren-Shaul, H., Boyeau, P., Kedmi, M., Pilzer, D., Jelinski, A., Yofe, I., David, E., Wagner, A., *et al.* (2022). DestVI identifies continuums of cell types in spatial
- 7596 transcriptomics data. Nat Biotechnol 40, 1360-1369.
- 7597
- 7598 Lorthongpanich, C., Cheow, L.F., Balu, S., Quake, S.R., Knowles, B.B., Burkholder, W.F., Solter, D. and Messerschmidt, D.M. (2013). Single-cell DNA-methylation analysis reveals
- 7599 epigenetic chimerism in preimplantation embryos. Science 341, 1110-1112.
- 7600
- 7601 Lotfollahi, M., Litinetskaya, A. and Theis, F.J. (2022). Multigrate: single-cell multi-omic data
- 7602 integration. bioRxiv, doi: 10.1101/2022.03.16.484643.
- 7603 Lovatt, D., Ruble, B.K., Lee, J., Dueck, H., Kim, T.K., Fisher, S., Francis, C., Spaethling, J.M.,
- 7604 Wolf, J.A., Grady, M.S., *et al.* (2014). Transcriptome in vivo analysis (TIVA) of spatially
- 7605 defined single cells in live tissue. Nat Methods 11, 190-196.
- 7606 Love, J.C., Ronan, J.L., Grotenbreg, G.M., van der Veen, A.G. and Ploegh, H.L. (2006). A
- 7607 microengraving method for rapid selection of single cells producing antigen-specific
- 7608 antibodies. Nat Biotechnol 24, 703-707.
- 7609 Love, M.I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion
- 7610 for RNA-seq data with DESeq2. Genome Biol 15, 550.
- 7611 Loza, M., Teraguchi, S., Standley, D.M. and Diez, D. (2022). Unbiased integration of single cell
- 7612 transcriptome replicates. NAR Genom Bioinform 4, lqac022.
- 7613 Lu, S., Conn, D.J., Chen, S., Johnson, K.D., Bresnick, E.H. and Keleş, S. (2021). MLG: multilayer
- 7614 graph clustering for multi-condition scRNA-seq data. Nucleic Acids Res 49, e127.
- 7615 Lubeck, E., Coskun, A.F., Zhiyentayev, T., Ahmad, M. and Cai, L. (2014). Single-cell *in situ* RNA
- 7616 profiling by sequential hybridization. Nat Methods 11, 360-361.
- 7617 Luecken, M.D., Buttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl,
- 7618 D.C., Zappia, L., Dugas, M., Colome-Tatche, M., *et al.* (2022). Benchmarking atlas-level
- 7619 data integration in single-cell genomics. Nat Methods 19, 41-50.
- 7620 Luecken, M.D. and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a
- 7621 tutorial. Mol Syst Biol 15, e8746.
- 7622 Lun, A.T.L., Riesenfeld, S., Andrews, T., Dao, T.P., Gomes, T. and Marioni, J.C. (2019).
- 7623 EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA
- 7624 sequencing data. Genome Biol 20, 63.
- 7625 Luo, C., Keown, C.L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J.R.,
- 7626 Sandoval, J.P., *et al.* (2017). Single-cell methylomes identify neuronal subtypes and
- 7627 regulatory elements in mammalian cortex. Science 357, 600-604.
- 7628 Lynch, A.W., Theodoris, C.V., Long, H.W., Brown, M., Liu, X.S. and Meyer, C.A. (2022). MIRA:
- 7629 joint regulatory modeling of multimodal expression and chromatin accessibility in single
- 7630 cells. Nat Methods 19, 1097-1108.
- 7631 Ma, A., Wang, X., Wang, C., Li, J., Xiao, T., Wang, J., Li, Y., Liu, Y., Chang, Y., Wang, D., *et al.*
- 7632 (2021a). DeepMAPS: Single-cell biological network inference using heterogeneous graph
- 7633 transformer. bioRxiv, doi: 10.1101/2021.10.31.466658.
- 7634 Ma, F. and Pellegrini, M. (2020). ACTINN: automated identification of cell types in single cell
- 7635 RNA sequencing. Bioinformatics 36, 533-538.
- 7636 Ma, Q., Li, S., Zhuang, W., Li, S., Wang, J. and Zeng, D. (2021b). Self-Supervised Time Series
- 7637 Clustering With Model-Based Dynamics. IEEE Trans Neural Netw Learn Syst 32,

- 7638 3942-3955.
- 7639 Ma, S., de la Fuente Revenga, M., Sun, Z., Sun, C., Murphy, T.W., Xie, H., González-Maeso, J.
7640 and Lu, C. (2018). Cell-type-specific brain methylomes profiled via ultralow-input
7641 microfluidics. *Nat Biomed Eng* 2, 183-194.
- 7642 Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K.,
7643 Tay, T., *et al.* (2020). Chromatin Potential Identified by Shared Single-Cell Profiling of RNA
7644 and Chromatin. *Cell* 183, 1103-1116.e1120.
- 7645 Ma, W., Su, K. and Wu, H. (2021c). Evaluation of some aspects in supervised cell type
7646 identification for single-cell RNA-seq: classifier, feature selection, and reference construction.
7647 *Genome Biol* 22, 264.
- 7648 Ma, Y. and Zhou, X. (2022). Spatially informed cell-type deconvolution for spatial transcriptomics.
7649 *Nat Biotechnol* 40, 1349-1359.
- 7650 Macaulay, I.C., Haerty, W., Kumar, P., Li, Y.I., Hu, T.X., Teng, M.J., Goolam, M., Saurat, N.,
7651 Coupland, P., Shirley, L.M., *et al.* (2015). G&T-seq: parallel sequencing of single-cell
7652 genomes and transcriptomes. *Nat Methods* 12, 519-522.
- 7653 Macaulay, I.C. and Voet, T. (2014). Single cell genomics: advances and future perspectives. *PLoS
7654 Genet* 10, e1004126.
- 7655 Macnair, W. and Robinson, M. (2023). SampleQC: robust multivariate, multi-cell type,
7656 multi-sample quality control for single-cell data. *Genome Biol* 24, 23.
- 7657 Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R.,
7658 Kamitaki, N., Martersteck, E.M., *et al.* (2015). Highly Parallel Genome-wide Expression
7659 Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161, 1202-1214.
- 7660 Macqueen, J. (1967). Some Methods for Classification and Analysis of MultiVariate Observations.
7661 In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability,
7662 281-297.
- 7663 Madisoone, E., Oliver, A.J., Kleshchevnikov, V., Wilbrey-Clark, A., Polanski, K., Orsi, A.R.,
7664 Mamanova, L., Bolt, L., Richoz, N., Elmentait, R., *et al.* (2021). A spatial multi-omics atlas
7665 of the human lung reveals a novel immune cell survival niche. *bioRxiv*, doi:
7666 10.1101/2021.11.26.470108.
- 7667 Mah, C.K., Ahmed, N., Lam, D., Monell, A., Kern, C., Han, Y., Cesnik, A.J., Lundberg, E., Zhu,
7668 Q., Carter, H., *et al.* (2022). Bento: A toolkit for subcellular analysis of spatial
7669 transcriptomics data. *bioRxiv*, doi: 10.1101/2022.06.10.495510.
- 7670 Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. and Church, G.M.
7671 (2013). RNA-guided human genome engineering via Cas9. *Science* 339, 823-826.
- 7672 Malta, T.M., Sokolov, A., Gentles, A.J., Burzykowski, T., Poisson, L., Weinstein, J.N., Kaminska,
7673 B., Huelsken, J., Omberg, L., Gevaert, O., *et al.* (2018). Machine Learning Identifies
7674 Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* 173, 338-354.e315.
- 7675 Maniatis, S., Aijo, T., Vickovic, S., Braine, C., Kang, K., Mollbrink, A., Fagegaltier, D.,
7676 Andrusivova, Z., Saarenpaa, S., Saiz-Castro, G., *et al.* (2019). Spatiotemporal dynamics of
7677 molecular pathology in amyotrophic lateral sclerosis. *Science* 364, 89-93.
- 7678 Marco, E., Karp, R.L., Guo, G., Robson, P., Hart, A.H., Trippa, L. and Yuan, G.C. (2014).
7679 Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc
7680 Natl Acad Sci U S A* 111, E5643-5650.
- 7681 Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M. and Wold, B.J.

- 7682 (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and
7683 RNA splicing. *Genome Res* 24, 496-510.
- 7684 Martin, P.C.N., Kim, H., Lovkvist, C., Hong, B.W. and Won, K.J. (2022). Vesalius: high-resolution
7685 in silico anatomization of spatial transcriptomic data using image analysis. *Mol Syst Biol* 18,
7686 e11080.
- 7687 Marx, V. (2019). A dream of single-cell proteomics. *Nat Methods* 16, 809-812.
- 7688 Maseda, F., Cang, Z. and Nie, Q. (2021). DEEPsc: A Deep Learning-Based Map Connecting
7689 Single-Cell Transcriptomics and Spatial Imaging Data. *Front Genet* 12, 636743.
- 7690 Masuda, T., Inamori, Y., Furukawa, A., Yamahiro, M., Momosaki, K., Chang, C.H., Kobayashi, D.,
7691 Ohguchi, H., Kawano, Y., Ito, S., *et al.* (2022). Water Droplet-in-Oil Digestion Method for
7692 Single-Cell Proteomics. *Anal Chem* 94, 10329-10336.
- 7693 McConnell, M.J., Lindberg, M.R., Brennand, K.J., Piper, J.C., Voet, T., Cowing-Zitron, C.,
7694 Shumilina, S., Lasken, R.S., Vermeesch, J.R., Hall, I.M., *et al.* (2013). Mosaic copy number
7695 variation in human neurons. *Science* 342, 632-637.
- 7696 McGinnis, C.S., Murrow, L.M. and Gartner, Z.J. (2019a). DoubletFinder: Doublet Detection in
7697 Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* 8,
7698 329-337.e324.
- 7699 McGinnis, C.S., Patterson, D.M., Winkler, J., Conrad, D.N., Hein, M.Y., Srivastava, V., Hu, J.L.,
7700 Murrow, L.M., Weissman, J.S., Werb, Z., *et al.* (2019b). MULTI-seq: sample multiplexing for
7701 single-cell RNA sequencing using lipid-tagged indices. *Nat Methods* 16, 619-626.
- 7702 McInnes, L. and Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for
7703 Dimension Reduction. *The Journal of Open Source Software* 3, 861.
- 7704 McPherson, A., Roth, A., Laks, E., Masud, T., Bashashati, A., Zhang, A.W., Ha, G., Biele, J., Yap,
7705 D., Wan, A., *et al.* (2016). Divergent modes of clonal spread and intraperitoneal mixing in
7706 high-grade serous ovarian cancer. *Nat Genet* 48, 758-767.
- 7707 Medaglia, C., Giladi, A., Stoler-Barak, L., De Giovanni, M., Salame, T.M., Biram, A., David, E.,
7708 Li, H., Iannaccone, M., Shulman, Z., *et al.* (2017). Spatial reconstruction of immune niches by
7709 combining photoactivatable reporters and scRNA-seq. *Science* 358, 1622-1626.
- 7710 Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., Zhu, M., Wu, J., Shi, X., Taing, L., *et al.*
7711 (2017). Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data
7712 in human and mouse. *Nucleic Acids Res* 45, D658-d662.
- 7713 Meier, F., Brunner, A.D., Koch, S., Koch, H., Lubeck, M., Krause, M., Goedecke, N., Decker, J.,
7714 Kosinski, T., Park, M.A., *et al.* (2018). Online Parallel Accumulation-Serial Fragmentation
7715 (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Mol Cell Proteomics* 17,
7716 2534-2545.
- 7717 Meir, Z., Mukamel, Z., Chomsky, E., Lifshitz, A. and Tanay, A. (2020). Single-cell analysis of
7718 clonal maintenance of transcriptional and epigenetic states in cancer cells. *Nat Genet* 52,
7719 709-718.
- 7720 Melsted, P., Booeshaghi, A.S., Liu, L., Gao, F., Lu, L., Min, K.H.J., da Veiga Beltrame, E.,
7721 Hjörleifsson, K.E., Gehring, J. and Pachter, L. (2021). Modular, efficient and
7722 constant-memory single-cell RNA-seq preprocessing. *Nat Biotechnol* 39, 813-818.
- 7723 Melsted, P., Ntranos, V. and Pachter, L. (2019). The barcode, UMI, set format and BUStools.
7724 *Bioinformatics* 35, 4472-4473.
- 7725 Meng, L., Wang, C., Shi, Y. and Luo, Q. (2021). Si-C is a method for inferring super-resolution

- 7726 intact genome structure from single-cell Hi-C data. *Nat Commun* 12, 4369.
- 7727 Merritt, C.R., Ong, G.T., Church, S.E., Barker, K., Danaher, P., Geiss, G., Hoang, M., Jung, J.,
7728 Liang, Y., McKay-Fleisch, J., *et al.* (2020). Multiplex digital spatial profiling of proteins and
7729 RNA in fixed tissue. *Nat Biotechnol* 38, 586-599.
- 7730 Meylan, M., Petitprez, F., Becht, E., Bougouin, A., Pupier, G., Calvez, A., Giglioli, I., Verkarre, V.,
7731 Lacroix, G., Verneau, J., *et al.* (2022). Tertiary lymphoid structures generate and propagate
7732 anti-tumor antibody-producing plasma cells in renal cell cancer. *Immunity* 55, 527-541.e525.
- 7733 Mezger, A., Klemm, S., Mann, I., Brower, K., Mir, A., Bostick, M., Farmer, A., Fordyce, P.,
7734 Linnarsson, S. and Greenleaf, W. (2018). High-throughput chromatin accessibility profiling
7735 at single-cell resolution. *Nat Commun* 9, 3647.
- 7736 Miao, Z., Deng, K., Wang, X. and Zhang, X. (2018). DEsingle for detecting three types of
7737 differential expression in single-cell RNA-seq data. *Bioinformatics* 34, 3223-3224.
- 7738 Michelsen, L., Reinders, M.J.T. and Mahfouz, A. (2021). Hierarchical progressive learning of cell
7739 identities in single-cell data. *Nat Commun* 12, 2799.
- 7740 Miller, B.F., Bambah-Mukku, D., Dulac, C., Zhuang, X. and Fan, J. (2021). Characterizing spatial
7741 gene expression heterogeneity in spatially resolved single-cell transcriptomic data with
7742 nonuniform cellular densities. *Genome Res* 31, 1843-1855.
- 7743 Miller, B.F., Huang, F., Atta, L., Sahoo, A. and Fan, J. (2022). Reference-free cell type
7744 deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nat
7745 Commun* 13, 2339.
- 7746 Miltenyi, S., Müller, W., Weichel, W. and Radbruch, A. (1990). High gradient magnetic cell
7747 separation with MACS. *Cytometry* 11, 231-238.
- 7748 Mimitou, E.P., Cheng, A., Montalbano, A., Hao, S., Stoeckius, M., Legut, M., Roush, T., Herrera,
7749 A., Papalexis, E., Ouyang, Z., *et al.* (2019). Multiplexed detection of proteins, transcriptomes,
7750 clonotypes and CRISPR perturbations in single cells. *Nat Methods* 16, 409-412.
- 7751 Minakshi, P., Ghosh, M., Kumar, R., Patki, H.S., Saini, H.M., Ranjan, K., Brar, B. and Prasad, G.
7752 Chapter 15 - Single-Cell Metabolomics: Technology and Applications. Academic Press 1,
7753 319-353.
- 7754 Ming, J., Lin, Z., Zhao, J., Wan, X., Yang, C. and Wu, A.R. (2022). FIRM: Flexible integration of
7755 single-cell RNA-sequencing data for large-scale multi-tissue cell atlas datasets. *Brief
7756 Bioinform* 23, bbac167.
- 7757 Minoura, K., Abe, K., Nam, H., Nishikawa, H. and Shimamura, T. (2021). A mixture-of-experts
7758 deep generative model for integrated analysis of single-cell multiomics data. *Cell Rep
7759 Methods* 1, 100071.
- 7760 Minussi, D.C., Nicholson, M.D., Ye, H., Davis, A., Wang, K., Baker, T., Tarabichi, M., Sei, E., Du,
7761 H., Rabbani, M., *et al.* (2021). Breast tumours maintain a reservoir of subclonal diversity
7762 during expansion. *Nature* 592, 302-308.
- 7763 Misra, B.B. (2020). Open-Source Software Tools, Databases, and Resources for Single-Cell and
7764 Single-Cell-Type Metabolomics. *Methods Mol Biol* 2064, 191-217.
- 7765 Mizuno, H., Tsuyama, N., Harada, T. and Masujima, T. (2008). Live single-cell video-mass
7766 spectrometry for cellular and subcellular molecular detection and cell classification. *J Mass
7767 Spectrom* 43, 1692-1700.
- 7768 Moehlin, J., Mollet, B., Colombo, B.M. and Mendoza-Parra, M.A. (2021). Inferring biologically
7769 relevant molecular tissue substructures by agglomerative clustering of digitized spatial

- 7770 transcriptomes with multilayer. *Cell Syst* 12, 694-705.e693.
- 7771 Moffitt, J.R., Bambah-Mukku, D., Eichhorn, S.W., Vaughn, E., Shekhar, K., Perez, J.D.,
7772 Rubinstein, N.D., Hao, J., Regev, A., Dulac, C., *et al.* (2018). Molecular, spatial, and
7773 functional single-cell profiling of the hypothalamic preoptic region. *Science* 362, eaau5324.
- 7774 Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J.C., Baron, M., Hajdu, C.H., Simeone,
7775 D.M. and Yanai, I. (2020). Integrating microarray-based spatial transcriptomics and
7776 single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat
7777 Biotechnol* 38, 333-342.
- 7778 Mooijman, D., Dey, S.S., Boisset, J.C., Crosetto, N. and van Oudenaarden, A. (2016). Single-cell
7779 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage
7780 reconstruction. *Nat Biotechnol* 34, 852-856.
- 7781 Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and
7782 quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5, 621-628.
- 7783 Moses, L. and Pachter, L. (2022). Museum of spatial transcriptomics. *Nat Methods* 19, 534-546.
- 7784 Mu, Q., Chen, Y. and Wang, J. (2019). Deciphering Brain Complexity Using Single-cell
7785 Sequencing. *Genomics Proteomics Bioinformatics* 17, 344-366.
- 7786 Mulqueen, R.M., Pokholok, D., Norberg, S.J., Torkenczy, K.A., Fields, A.J., Sun, D., Sinnamon,
7787 J.R., Shendure, J., Trapnell, C., O'Roak, B.J., *et al.* (2018). Highly scalable generation of
7788 DNA methylation profiles in single cells. *Nat Biotechnol* 36, 428-431.
- 7789 Mund, A., Brunner, A.-D. and Mann, M. (2022a). Unbiased spatial proteomics with single-cell
7790 resolution in tissues. *Mol Cell* 82, 2335-2349.
- 7791 Mund, A., Coscia, F., Kriston, A., Hollandi, R., Kovács, F., Brunner, A.-D., Migh, E., Schweizer,
7792 L., Santos, A., Bzorek, M., *et al.* (2022b). Deep Visual Proteomics defines single-cell identity
7793 and heterogeneity. *Nat Biotechnol* 40, 1231-1240.
- 7794 Muntel, J., Kirkpatrick, J., Bruderer, R., Huang, T., Vitek, O., Ori, A. and Reiter, L. (2019).
7795 Comparison of Protein Quantification in a Complex Background by DIA and TMT
7796 Workflows with Fixed Instrument Time. *J Proteome Res* 18, 1340-1351.
- 7797 Muskovic, W. and Powell, J.E. (2021). DropletQC: improved identification of empty droplets and
7798 damaged cells in single-cell RNA-seq data. *Genome Biol* 22, 329.
- 7799 Muto, Y., Wilson, P.C., Ledru, N., Wu, H., Dimke, H., Waikar, S.S. and Humphreys, B.D. (2021).
7800 Single cell transcriptional and chromatin accessibility profiling redefine cellular
7801 heterogeneity in the adult human kidney. *Nat Commun* 12, 2190.
- 7802 Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A.
7803 and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome
7804 structure. *Nature* 502, 59-64.
- 7805 Narayan, A., Berger, B. and Cho, H. (2021). Assessing single-cell transcriptomic variability
7806 through density-preserving data visualization. *Nat Biotechnol* 39, 765-774.
- 7807 Nault, R., Saha, S., Bhattacharya, S., Dodson, J., Sinha, S., Maiti, T. and Zacharewski, T. (2022).
7808 Benchmarking of a Bayesian single cell RNAseq differential gene expression test for
7809 dose-response study designs. *Nucleic Acids Res* 50, e48.
- 7810 Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A.,
7811 Levy, D., Esposito, D., *et al.* (2011). Tumour evolution inferred by single-cell sequencing.
7812 *Nature* 472, 90-94.
- 7813 Nehar-Belaid, D., Hong, S., Marches, R., Chen, G., Bolisetty, M., Baisch, J., Walters, L., Punaro,

- 7814 M., Rossi, R.J., Chung, C.H., *et al.* (2020). Mapping systemic lupus erythematosus
 7815 heterogeneity at the single-cell level. *Nat Immunol* 21, 1094-1106.
- 7816 Nemes, P. and Vertes, A. (2007). Laser ablation electrospray ionization for atmospheric pressure,
 7817 in vivo, and imaging mass spectrometry. *Anal Chem* 79, 8098-8106.
- 7818 Nestorowa, S., Hamey, F.K., Pijuan Sala, B., Diamanti, E., Shepherd, M., Laurenti, E., Wilson,
 7819 N.K., Kent, D.G. and Göttgens, B. (2016). A single-cell resolution map of mouse
 7820 hematopoietic stem and progenitor cell differentiation. *Blood* 128, e20-31.
- 7821 Network, C.G.A. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*
 7822 490, 61-70.
- 7823 Nguyen, S.N., Liyu, A.V., Chu, R.K., Anderton, C.R. and Laskin, J. (2017). Constant-Distance
 7824 Mode Nanospray Desorption Electrospray Ionization Mass Spectrometry Imaging of
 7825 Biological Samples with Complex Topography. *Anal Chem* 89, 1131-1137.
- 7826 Ni, X., Zhuo, M., Su, Z., Duan, J., Gao, Y., Wang, Z., Zong, C., Bai, H., Chapman, A.R., Zhao, J.,
 7827 *et al.* (2013). Reproducible copy number variation patterns among single circulating tumor
 7828 cells of lung cancer patients. *Proc Natl Acad Sci U.S.A* 110, 21083-21088.
- 7829 Ni, Z., Prasad, A., Chen, S., Halberg, R.B., Arkin, L.M., Drolet, B.A., Newton, M.A. and
 7830 Kendziora, C. (2022). SpotClean adjusts for spot swapping in spatial transcriptomics data.
 7831 *Nat Commun* 13, 2971.
- 7832 Nicin, L., Abplanalp, W.T., Schärzer, A., Sprengel, A., John, D., Mellentin, H., Tombor, L.,
 7833 Keuper, M., Ullrich, E., Klingel, K., *et al.* (2021). Single Nuclei Sequencing Reveals Novel
 7834 Insights Into the Regulation of Cellular Signatures in Children With Dilated Cardiomyopathy.
 7835 *Circulation* 143, 1704-1719.
- 7836 Niebler, S., Müller, A., Hankeln, T. and Schmidt, B. (2020). RainDrop: Rapid activation matrix
 7837 computation for droplet-based single-cell RNA-seq reads. *BMC Bioinformatics* 21, 274.
- 7838 Nitzan, M., Karaiskos, N., Friedman, N. and Rajewsky, N. (2019). Gene expression cartography.
 7839 *Nature* 576, 132-137.
- 7840 Noël, F., Massenet-Regad, L., Carmi-Levy, I., Cappuccio, A., Grandclaudon, M., Trichot, C.,
 7841 Kieffer, Y., Mechta-Grigoriou, F. and Soumelis, V. (2021). Dissection of intercellular
 7842 communication using the transcriptome-based framework ICELLNET. *Nat Commun* 12,
 7843 1089.
- 7844 Nwosu, A.J., Misal, S.A., Truong, T., Carson, R.H., Webber, K.G.I., Axtell, N.B., Liang, Y.,
 7845 Johnston, S.M., Virgin, K.L., Smith, E.G., *et al.* (2022). In-Depth Mass Spectrometry-Based
 7846 Proteomics of Formalin-Fixed, Paraffin-Embedded Tissues with a Spatial Resolution of
 7847 50-200 µm. *J Proteome Res* 21, 2237-2245.
- 7848 Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999). KEGG: Kyoto
 7849 Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 27, 29-34.
- 7850 Ortiz, C., Navarro, J.F., Jurek, A., Märtin, A., Lundeberg, J. and Meletis, K. (2020). Molecular
 7851 atlas of the adult mouse brain. *Sci Adv* 6, eabb3446.
- 7852 Padron-Manrique, C., Vázquez-Jiménez, A., Esquivel-Hernandez, D.A., Lopez, Y.E.M.,
 7853 Neri-Rosario, D., Sánchez-Castañeda, J.P., Giron-Villalobos, D. and Resendis-Antonio, O.
 7854 (2022). Diffusion on PCA-UMAP manifold captures a well-balance of local, global, and
 7855 continuum structure to denoise single-cell RNA sequencing data. *bioRxiv*, doi:
 7856 10.1101/2022.06.09.495525.
- 7857 Paik, D.T., Tian, L., Williams, I.M., Rhee, S., Zhang, H., Liu, C., Mishra, R., Wu, S.M.,

- 7858 Red-Horse, K. and Wu, J.C. (2020). Single-Cell RNA Sequencing Unveils Unique
 7859 Transcriptomic Signatures of Organ-Specific Endothelial Cells. *Circulation* 142, 1848-1862.
- 7860 Pan, N., Rao, W., Kothapalli, N.R., Liu, R., Burgett, A.W. and Yang, Z. (2014). The single-probe:
 7861 a miniaturized multifunctional device for single cell mass spectrometry analysis. *Anal Chem*
 7862 86, 9376-9380.
- 7863 Pan, N., Rao, W. and Yang, Z. (2020). Single-Probe Mass Spectrometry Analysis of Metabolites in
 7864 Single Cells. *Methods Mol Biol* 2064, 61-71.
- 7865 Pan, N., Standke, S.J., Kothapalli, N.R., Sun, M., Bensen, R.C., Burgett, A.W.G. and Yang, Z.
 7866 (2019). Quantification of Drug Molecules in Live Single Cells Using the Single-Probe Mass
 7867 Spectrometry Technique. *Anal Chem* 91, 9018-9024.
- 7868 Pang, M., Su, K. and Li, M. (2021). Leveraging information in spatial transcriptomics to predict
 7869 super-resolution gene expression from histology images in tumors. *bioRxiv*, doi:
 7870 10.1101/2021.11.28.470212.
- 7871 Papalexi, E., Mimitou, E.P., Butler, A.W., Foster, S., Bracken, B., Mauck, W.M., 3rd, Wessels,
 7872 H.H., Hao, Y., Yeung, B.Z., Smibert, P., *et al.* (2021). Characterizing the molecular regulation
 7873 of inhibitory immune checkpoints with multimodal single-cell screens. *Nat Genet* 53,
 7874 322-331.
- 7875 Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. and Hellmann, I. (2018). zUMIs - A fast and
 7876 flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* 7, giy059.
- 7877 Park, J., Choi, W., Tiesmeyer, S., Long, B., Borm, L.E., Garren, E., Nguyen, T.N., Tasic, B.,
 7878 Codeluppi, S., Graf, T., *et al.* (2021). Cell segmentation-free inference of cell types from in
 7879 situ transcriptomics data. *Nat Commun* 12, 3545.
- 7880 Passarelli, M.K., Pirkl, A., Moellers, R., Grinfeld, D., Kollmer, F., Havelund, R., Newman, C.F.,
 7881 Marshall, P.S., Arlinghaus, H., Alexander, M.R., *et al.* (2017). The 3D OrbiSIMS-label-free
 7882 metabolic imaging with subcellular lateral resolution and high mass-resolving power. *Nat
 7883 Methods* 14, 1175-1183.
- 7884 Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. and Kingsford, C. (2017). Salmon provides fast
 7885 and bias-aware quantification of transcript expression. *Nat Methods* 14, 417-419.
- 7886 Patro, R., Mount, S.M. and Kingsford, C. (2014). Sailfish enables alignment-free isoform
 7887 quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol* 32,
 7888 462-464.
- 7889 Patty, B.J. and Hainer, S.J. (2021). Transcription factor chromatin profiling genome-wide using
 7890 uliCUT&RUN in single cells and individual blastocysts. *Nat Protoc* 16, 2633-2666.
- 7891 Pei, G., Yan, F., Simon, L.M., Dai, Y., Jia, P. and Zhao, Z. (2023). deCS: A Tool for Systematic
 7892 Cell Type Annotations of Single-cell RNA Sequencing Data among Human Tissues.
 7893 *Genomics Proteomics Bioinformatics* 21, 370-384.
- 7894 Peng, G., Suo, S., Cui, G., Yu, F., Wang, R., Chen, J., Chen, S., Liu, Z., Chen, G., Qian, Y., *et al.*
 7895 (2019). Molecular architecture of lineage allocation and tissue organization in early mouse
 7896 embryo. *Nature* 572, 528-532.
- 7897 Peng, M., Wamsley, B., Elkins, A.G., Geschwind, D.H., Wei, Y. and Roeder, K. (2021a). Cell type
 7898 hierarchy reconstruction via reconciliation of multi-resolution cluster tree. *Nucleic Acids Res*
 7899 49, e91.
- 7900 Peng, T., Chen, G.M. and Tan, K. (2021b). GLUER: integrative analysis of single-cell omics and
 7901 imaging data by deep neural network. *bioRxiv*, doi: 10.1101/2021.01.25.427845.

- 7902 Peres-Neto, P.R., Jackson, D.A. and Somers, K.M. (2005). How many principal components?
 7903 stopping rules for determining the number of non-trivial axes revisited. Computational
 7904 Statistics and Data Analysis 49, 974-997.
- 7905 Peterson, V.M., Zhang, K.X., Kumar, N., Wong, J., Li, L., Wilson, D.C., Moore, R., McClanahan,
 7906 T.K., Sadekova, S. and Klappenbach, J.A. (2017). Multiplexed quantification of proteins and
 7907 transcripts in single cells. Nat Biotechnol 35, 936-939.
- 7908 Petrany, M.J., Swoboda, C.O., Sun, C., Chetal, K., Chen, X., Weirauch, M.T., Salomonis, N. and
 7909 Millay, D.P. (2020). Single-nucleus RNA-seq identifies transcriptional heterogeneity in
 7910 multinucleated skeletal myofibers. Nat Commun 11, 6374.
- 7911 Petukhov, V., Guo, J., Baryawno, N., Severe, N., Scadden, D.T., Samsonova, M.G. and
 7912 Kharchenko, P.V. (2018). dropEst: pipeline for accurate estimation of molecular counts in
 7913 droplet-based single-cell RNA-seq experiments. Genome Biol 19, 78.
- 7914 Petukhov, V., Xu, R.J., Soldatov, R.A., Cadinu, P., Khodosevich, K., Moffitt, J.R. and Kharchenko,
 7915 P.V. (2022). Cell segmentation in imaging-based spatial transcriptomics. Nat Biotechnol 40,
 7916 345-354.
- 7917 Pham, D., Tan, X., Xu, J., Grice, L.F., Lam, P.Y., Raghubar, A., Vukovic, J., Ruitenberg, M.J. and
 7918 Nguyen, Q. (2020). stLearn: integrating spatial location, tissue morphology and gene
 7919 expression to find cell types, cell-cell interactions and spatial trajectories within
 7920 undissociated tissues. bioRxiv, doi: 10.1101/2020.05.31.125658.
- 7921 Picelli, S. (2017). Single-cell RNA-sequencing: The future of genome biology is now. RNA Biol
 7922 14, 637-650.
- 7923 Picelli, S., Björklund Å, K., Faridani, O.R., Sagasser, S., Winberg, G. and Sandberg, R. (2013).
 7924 Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods 10,
 7925 1096-1098.
- 7926 Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S. and Sandberg, R. (2014).
 7927 Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc 9, 171-181.
- 7928 Picher Á, J., Budeus, B., Wafzig, O., Krüger, C., Garc ía-G ómez, S., Mart ínez-Jim énez, M.I.,
 7929 D áz-Talavera, A., Weber, D., Blanco, L. and Schneider, A. (2016). TruePrime is a novel
 7930 method for whole-genome amplification from single cells based on TthPrimPol. Nat
 7931 Commun 7, 13296.
- 7932 Pierce, S.E., Granja, J.M. and Greenleaf, W.J. (2021). High-throughput single-cell chromatin
 7933 accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer.
 7934 Nat Commun 12, 2969.
- 7935 Pieters, R., De Lorenzo, P., Ancliffe, P., Aversa, L.A., Brethon, B., Biondi, A., Campbell, M.,
 7936 Escherich, G., Ferster, A., Gardner, R.A., *et al.* (2019). Outcome of Infants Younger Than 1
 7937 Year With Acute Lymphoblastic Leukemia Treated With the Interfant-06 Protocol: Results
 7938 From an International Phase III Randomized Study. J Clin Oncol 37, 2246-2256.
- 7939 Pliner, H.A., Packer, J.S., McFalone-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie,
 7940 D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., *et al.* (2018). Cicero Predicts
 7941 cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol Cell
 7942 71, 858-871.e858.
- 7943 Pliner, H.A., Shendure, J. and Trapnell, C. (2019). Supervised classification enables rapid
 7944 annotation of cell atlases. Nat Methods 16, 983-986.
- 7945 Polański, K., Young, M.D., Miao, Z., Meyer, K.B., Teichmann, S.A. and Park, J.E. (2020).

- 7946 BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* 36, 964-965.
- 7947 Prabhakaran, S. (2022). Sparcle: assigning transcripts to cells in multiplexed images. *Bioinform*
7948 *Adv* 2, vbac048.
- 7949 Preissl, S., Fang, R., Huang, H., Zhao, Y., Raviram, R., Gorkin, D.U., Zhang, Y., Sos, B.C., Afzal,
7950 V., Dickel, D.E., *et al.* (2018). Single-nucleus analysis of accessible chromatin in developing
7951 mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat Neurosci* 21,
7952 432-439.
- 7953 Prieto-Vila, M., Usuba, W., Takahashi, R.U., Shimomura, I., Sasaki, H., Ochiya, T. and Yamamoto,
7954 Y. (2019). Single-Cell Analysis Reveals a Preexisting Drug-Resistant Subpopulation in the
7955 Luminal Breast Cancer Subtype. *Cancer Res* 79, 4412-4425.
- 7956 Przytycki, P.F. and Pollard, K.S. (2022). CellWalkR: an R package for integrating and visualizing
7957 single-cell and bulk data to resolve regulatory elements. *Bioinformatics* 38, 2621-2623.
- 7958 Qi, J., Sun, H., Zhang, Y., Wang, Z., Xun, Z., Li, Z., Ding, X., Bao, R., Hong, L., Jia, W., *et al.*
7959 (2022). Single-cell and spatial analysis reveal interaction of FAP(+) fibroblasts and SPP1(+)
7960 macrophages in colorectal cancer. *Nat Commun* 13, 1742.
- 7961 Qi, R., Wu, J., Guo, F., Xu, L. and Zou, Q. (2021). A spectral clustering with self-weighted
7962 multiple kernel learning method for single-cell RNA-seq data. *Brief Bioinform* 22, bbaa216.
- 7963 Qian, X., Harris, K.D., Hauling, T., Nicoloutsopoulos, D., Munoz-Manchado, A.B., Skene, N.,
7964 Hjerling-Leffler, J. and Nilsson, M. (2020). Probabilistic cell typing enables fine mapping of
7965 closely related cell types *in situ*. *Nat Methods* 17, 101-106.
- 7966 Qiao, C. and Huang, Y. (2021). Representation learning of RNA velocity reveals robust cell
7967 transitions. *Proc Natl Acad Sci U S A* 118, e2105859118.
- 7968 Qiu, P., Simonds, E.F., Bendall, S.C., Gibbs, K.D., Jr., Bruggner, R.V., Linderman, M.D., Sachs,
7969 K., Nolan, G.P. and Plevritis, S.K. (2011). Extracting a cellular hierarchy from
7970 high-dimensional cytometry data with SPADE. *Nat Biotechnol* 29, 886-891.
- 7971 Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A. and Trapnell, C. (2017). Reversed
7972 graph embedding resolves complex single-cell trajectories. *Nat Methods* 14, 979-982.
- 7973 Ramani, V., Deng, X., Qiu, R., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S.,
7974 Duan, Z. and Shendure, J. (2017). Massively multiplex single-cell Hi-C. *Nat Methods* 14,
7975 263-266.
- 7976 Ramilowski, J.A., Goldberg, T., Harshbarger, J., Kloppmann, E., Lizio, M., Satagopam, V.P., Itoh,
7977 M., Kawaji, H., Carninci, P., Rost, B., *et al.* (2015). A draft network of
7978 ligand-receptor-mediated multicellular signalling in human. *Nat Commun* 6, 7866.
- 7979 Ramsköld, D., Luo, S., Wang, Y.C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova,
7980 I., Loring, J.F., Laurent, L.C., *et al.* (2012). Full-length mRNA-Seq from single-cell levels of
7981 RNA and individual circulating tumor cells. *Nat Biotechnol* 30, 777-782.
- 7982 Ran, D., Zhang, S., Lytal, N. and An, L. (2020). scDoc: correcting drop-out events in single-cell
7983 RNA-seq data. *Bioinformatics* 36, 4233-4239.
- 7984 Rao, W., Pan, N. and Yang, Z. (2015). High Resolution Tissue Imaging Using the Single-probe
7985 Mass Spectrometry under Ambient Conditions. *J Am Soc Mass Spectrom* 26, 986-993.
- 7986 Rappez, L., Stadler, M., Triana, S., Gathungu, R.M., Ovchinnikova, K., Phapale, P., Heikenwalder,
7987 M. and Alexandrov, T. (2021). SpaceM reveals metabolic states of single cells. *Nat Methods*
7988 18, 799-805.
- 7989 Reed, E.R. and Monti, S. (2021). Multi-resolution characterization of molecular taxonomies in

- 7990 bulk and single-cell transcriptomics data. Nucleic Acids Res 49, e98.
- 7991 Ren, H., Walker, B.L., Cang, Z. and Nie, Q. (2022). Identifying multicellular spatiotemporal
7992 organization of cells with SpaceFlow. Nat Commun 13, 4076.
- 7993 Replogle, J.M., Norman, T.M., Xu, A., Hussmann, J.A., Chen, J., Cogan, J.Z., Meer, E.J., Terry,
7994 J.M., Riordan, D.P., Srinivas, N., *et al.* (2020). Combinatorial single-cell CRISPR screens by
7995 direct guide RNA capture and targeted sequencing. Nat Biotechnol 38, 954-961.
- 7996 Replogle, J.M., Saunders, R.A., Pogson, A.N., Hussmann, J.A., Lenail, A., Guna, A., Mascibroda,
7997 L., Wagner, E.J., Adelman, K., Lithwick-Yanai, G., *et al.* (2022). Mapping information-rich
7998 genotype-phenotype landscapes with genome-scale Perturb-seq. Cell 185, 2559-2575.e2528.
- 7999 Riba, A., Oravecz, A., Durik, M., Jiménez, S., Alunni, V., Cerciat, M., Jung, M., Keime, C., Keyes,
8000 W.M. and Molina, N. (2022). Cell cycle gene regulation dynamics revealed by RNA velocity
8001 and deep-learning. Nat Commun 13, 2865.
- 8002 Richman, L.P., Goyal, Y., Jiang, C.L. and Raj, A. (2023). ClonoCluster: A method for using clonal
8003 origin to inform transcriptome clustering. Cell Genom 3, 100247.
- 8004 Riebensahm, C., Joosse, S.A., Mohme, M., Hanssen, A., Matschke, J., Goy, Y., Witzel, I., Lamszus,
8005 K., Kropidlowski, J., Petersen, C., *et al.* (2019). Clonality of circulating tumor cells in breast
8006 cancer brain metastasis patients. Breast Cancer Res 21, 101.
- 8007 Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. and Vert, J.P. (2018). A general and flexible
8008 method for signal extraction from single-cell RNA-seq data. Nat Commun 9, 284.
- 8009 Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015). limma
8010 powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic
8011 Acids Res 43, e47.
- 8012 Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010). edgeR: a Bioconductor package for
8013 differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140.
- 8014 Robinson, M.D. and Oshlack, A. (2010). A scaling normalization method for differential
8015 expression analysis of RNA-seq data. Genome Biol 11, R25.
- 8016 Rodriguez-Meira, A., Buck, G., Clark, S.A., Povinelli, B.J., Alcolea, V., Louka, E., McGowan, S.,
8017 Hamblin, A., Sousos, N., Barkas, N., *et al.* (2019). Unravelling Intratumoral Heterogeneity
8018 through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. Mol
8019 Cell 73, 1292-1305.e1298.
- 8020 Rodrigues, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J.,
8021 Chen, L.M., Chen, F. and Macosko, E.Z. (2019). Slide-seq: A scalable technology for
8022 measuring genome-wide expression at high spatial resolution. Science 363, 1463-1467.
- 8023 Roohani, Y., Huang, K. and Leskovec, J. (2023). Predicting transcriptional outcomes of novel
8024 multigene perturbations with GEARS. Nat Biotechnol, doi: 10.1038/s41587-023-01905-6.
- 8025 Rooijers, K., Markodimitraki, C.M., Rang, F.J., de Vries, S.S., Chialastri, A., de Luca, K.L.,
8026 Mooijman, D., Dey, S.S. and Kind, J. (2019). Simultaneous quantification of protein-DNA
8027 contacts and transcriptomes in single cells. Nat Biotechnol 37, 766-772.
- 8028 Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T.,
8029 Peeler, D.J., Mukherjee, S., Chen, W., *et al.* (2018). Single-cell profiling of the developing
8030 mouse brain and spinal cord with split-pool barcoding. Science 360, 176-182.
- 8031 Rosenthal, M., Bryner, D., Huffer, F., Evans, S., Srivastava, A. and Neretti, N. (2019). Bayesian
8032 Estimation of Three-Dimensional Chromosomal Structure from Single-Cell Hi-C Data. J
8033 Comput Biol 26, 1191-1202.

- 8034 Rotem, A., Ram, O., Shores, N., Sperling, R.A., Goren, A., Weitz, D.A. and Bernstein, B.E.
 8035 (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* 33, 1165-1172.
- 8036
- 8037 Roth, T.L., Li, P.J., Blaesche, F., Nies, J.F., Apathy, R., Mowery, C., Yu, R., Nguyen, M.L.T., Lee,
 8038 Y., Truong, A., *et al.* (2020). Pooled Knockin Targeting for Genome Engineering of Cellular
 8039 Immunotherapies. *Cell* 181, 728-744.e721.
- 8040 Rubakhin, S.S., Lanni, E.J. and Sweedler, J.V. (2013). Progress toward single cell metabolomics.
 8041 *Current Opin Biotechnol* 24, 95-104.
- 8042 Rubin, A.J., Parker, K.R., Satpathy, A.T., Qi, Y., Wu, B., Ong, A.J., Mumbach, M.R., Ji, A.L., Kim,
 8043 D.S., Cho, S.W., *et al.* (2019). Coupled Single-Cell CRISPR Screening and Epigenomic
 8044 Profiling Reveals Causal Gene Regulatory Networks. *Cell* 176, 361-376.e317.
- 8045 Rubio, C., Rodrigo, L., Mir, P., Mateu, E., Peinado, V., Milán, M., Al-Asmar, N., Campos-Galindo,
 8046 I., García, S. and Simón, C. (2013). Use of array comparative genomic hybridization
 8047 (array-CGH) for embryo assessment: clinical results. *Fertil Steril* 99, 1044-1048.
- 8048 Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G.,
 8049 Montrone, C. and Mewes, H.W. (2010). CORUM: the comprehensive resource of
 8050 mammalian protein complexes--2009. *Nucleic Acids Res* 38, D497-501.
- 8051 Santinha, A.J., Klingler, E., Kuhn, M., Farouni, R., Lagler, S., Kalamakis, G., Lischetti, U.,
 8052 Jabaoudon, D. and Platt, R.J. (2023). Transcriptional linkage analysis with *in vivo*
 8053 AAV-Perturb-seq. *Nature* 622, 367-375.
- 8054 Sasagawa, Y., Danno, H., Takada, H., Ebisawa, M., Tanaka, K., Hayashi, T., Kurisaki, A. and
 8055 Nikaido, I. (2018). Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that
 8056 effectively uses limited sequence reads. *Genome Biol* 19, 29.
- 8057 Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K.D., Imai, T. and Ueda, H.R. (2013).
 8058 Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals
 8059 non-genetic gene-expression heterogeneity. *Genome Biol* 14, R31.
- 8060 Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. and Regev, A. (2015). Spatial reconstruction of
 8061 single-cell gene expression data. *Nat Biotechnol* 33, 495-502.
- 8062 Satpathy, A.T., Granja, J.M., Yost, K.E., Qi, Y., Meschi, F., McDermott, G.P., Olsen, B.N.,
 8063 Mumbach, M.R., Pierce, S.E., Corces, M.R., *et al.* (2019). Massively parallel single-cell
 8064 chromatin landscapes of human immune cell development and intratumoral T cell exhaustion.
 8065 *Nat Biotechnol* 37, 925-936.
- 8066 Savas, P., Virassamy, B., Ye, C., Salim, A., Mintoff, C.P., Caramia, F., Salgado, R., Byrne, D.J.,
 8067 Teo, Z.L., Dushyanthen, S., *et al.* (2018). Single-cell profiling of breast cancer T cells reveals
 8068 a tissue-resident memory subset associated with improved prognosis. *Nat Med* 24, 986-993.
- 8069 Schatz, M.C. (2009). CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics*
 8070 25, 1363-1369.
- 8071 Schede, H.H., Schneider, C.G., Stergiadou, J., Borm, L.E., Ranjak, A., Yamawaki, T.M., David,
 8072 F.P.A., Lönnberg, P., Tosches, M.A., Codeluppi, S., *et al.* (2021). Spatial tissue profiling by
 8073 imaging-free molecular tomography. *Nat Biotechnol* 39, 968-977.
- 8074 Schep, A.N., Wu, B., Buenrostro, J.D. and Greenleaf, W.J. (2017). chromVAR: inferring
 8075 transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods*
 8076 14, 975-978.
- 8077 Schoof, E.M., Furtwängler, B., Üresin, N., Rapin, N., Savickas, S., Gentil, C., Lechman, E., Keller,

- 8078 U.A.D., Dick, J.E. and Porse, B.T. (2021). Quantitative single-cell proteomics as a tool to
 8079 characterize cellular hierarchies. *Nat Commun* 12, 3341.
- 8080 Schubert, S.M., Walter, S.R., Manesse, M. and Walt, D.R. (2016). Protein Counting in Single
 8081 Cancer Cells. *Anal Chem* 88, 2952-2957.
- 8082 Scialdone, A., Tanaka, Y., Jawaid, W., Moignard, V., Wilson, N.K., Macaulay, I.C., Marioni, J.C.
 8083 and Göttgens, B. (2016). Resolving early mesoderm diversification through single-cell
 8084 expression profiling. *Nature* 535, 289-293.
- 8085 See, P., Lum, J., Chen, J. and Ginhoux, F. (2018). A Single-Cell Sequencing Guide for
 8086 Immunologists. *Front Immunol* 9, 2425.
- 8087 Senabouth, A., Lukowski, S.W., Hernandez, J.A., Andersen, S.B., Mei, X., Nguyen, Q.H. and
 8088 Powell, J.E. (2019). ascend: R package for analysis of single-cell RNA-seq data. *Gigascience*
 8089 8, giz087.
- 8090 Sethi, A., Gu, M., Gumusgoz, E., Chan, L., Yan, K.K., Rozowsky, J., Barozzi, I., Afzal, V.,
 8091 Akiyama, J.A., Plajzer-Frick, I., *et al.* (2020). Supervised enhancer prediction with epigenetic
 8092 pattern recognition and targeted validation. *Nat Methods* 17, 807-814.
- 8093 Setty, M., Tadmor, M.D., Reich-Zeliger, S., Angel, O., Salame, T.M., Kathail, P., Choi, K., Bendall,
 8094 S., Friedman, N. and Pe'er, D. (2016). Wishbone identifies bifurcating developmental
 8095 trajectories from single-cell data. *Nat Biotechnol* 34, 637-645.
- 8096 Shah, S., Lubeck, E., Zhou, W. and Cai, L. (2016). In Situ Transcription Profiling of Single Cells
 8097 Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* 92, 342-357.
- 8098 Shahryary, Y., Hazarika, R.R. and Johannes, F. (2020). MethylStar: A fast and robust
 8099 pre-processing pipeline for bulk or single-cell whole-genome bisulfite sequencing data. *BMC*
 8100 *Genomics* 21, 479.
- 8101 Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T., Heckl, D., Ebert,
 8102 B.L., Root, D.E., Doench, J.G., *et al.* (2014). Genome-scale CRISPR-Cas9 knockout
 8103 screening in human cells. *Science* 343, 84-87.
- 8104 Shang, W., Zhang, Y., Shu, M., Wang, W., Ren, L., Chen, F., Shao, L., Lu, S., Bo, S., Ma, S., *et al.*
 8105 (2018). Comprehensive chromosomal and mitochondrial copy number profiling in human
 8106 IVF embryos. *Reprod Biomed Online* 36, 67-74.
- 8107 Shao, X., Li, C., Yang, H., Lu, X., Liao, J., Qian, J., Wang, K., Cheng, J., Yang, P., Chen, H., *et al.*
 8108 (2022a). Knowledge-graph-based cell-cell communication inference for spatially resolved
 8109 transcriptomic data with SpaTalk. *Nat Commun* 13, 4429.
- 8110 Shao, X., Liao, J., Li, C., Lu, X., Cheng, J. and Fan, X. (2021a). CellTalkDB: a manually curated
 8111 database of ligand-receptor interactions in humans and mice. *Brief Bioinform* 22.
- 8112 Shao, X., Liao, J., Lu, X., Xue, R., Ai, N. and Fan, X. (2020a). scCATCH: Automatic Annotation
 8113 on Cell Types of Clusters from Single-Cell RNA Sequencing Data. *iScience* 23, 100882.
- 8114 Shao, X., Lu, X., Liao, J., Chen, H. and Fan, X. (2020b). New avenues for systematically inferring
 8115 cell-cell communication: through single-cell transcriptomics data. *Protein Cell* 11, 866-880.
- 8116 Shao, X., Wang, X., Guan, S., Lin, H., Yan, G., Gao, M., Deng, C. and Zhang, X. (2018).
 8117 Integrated Proteome Analysis Device for Fast Single-Cell Protein Profiling. *Anal Chem* 90,
 8118 14003-14010.
- 8119 Shao, X., Wang, Z., Wang, K., Lu, X., Zhang, P., Guo, R., Liao, J., Yang, P., Xu, X. and Fan, X.
 8120 (2024). A Single-Cell Landscape of Human Liver Transplantation Reveals a Pathogenic
 8121 Immune Niche Associated with Early Allograft Dysfunction. *Engineering*, doi:

- 8122 10.1016/j.eng.2023.12.004.
- 8123 Shao, X., Yang, H., Zhuang, X., Liao, J., Yang, P., Cheng, J., Lu, X., Chen, H. and Fan, X. (2021b).
8124 scDeepSort: a pre-trained cell-type annotation method for single-cell transcriptomics using
8125 deep learning with a weighted graph neural network. Nucleic Acids Res 49, e122.
- 8126 Shao, Y., Zhou, Y., Liu, Y., Zhang, W., Zhu, G., Zhao, Y., Zhang, Q., Yao, H., Zhao, H., Guo, G., *et*
8127 *al.* (2022b). Intact living-cell electrolaunching ionization mass spectrometry for single-cell
8128 metabolomics. Chem Sci 13, 8065-8073.
- 8129 Shareef, S.J., Bevill, S.M., Raman, A.T., Aryee, M.J., van Galen, P., Hovestadt, V. and Bernstein,
8130 B.E. (2021). Extended-representation bisulfite sequencing of gene regulatory elements in
8131 multiplexed samples and single cells. Nat Biotechnol 39, 1086-1094.
- 8132 Shen, Z., Zhang, R., Huang, Y., Chen, J., Yu, M., Li, C., Zhang, Y., Chen, L., Huang, X., Yang, J.,
8133 *et al.* (2024). The spatial transcriptomic landscape of human gingiva in health and
8134 periodontitis. Sci China Life Sci 67, 720-732.
- 8135 Sheng, K., Cao, W., Niu, Y., Deng, Q. and Zong, C. (2017). Effective detection of variation in
8136 single-cell transcriptomes using MATQ-seq. Nat Methods 14, 267-270.
- 8137 Shomroni, O., Sitte, M., Schmidt, J., Parbin, S., Ludewig, F., Yigit, G., Zelarayan, L.C.,
8138 Streckfuss-B ömeke, K., Wollnik, B. and Salinas, G. (2022). A novel single-cell
8139 RNA-sequencing approach and its applicability connecting genotype to phenotype in ageing
8140 disease. Sci Rep 12, 4091.
- 8141 Shrestha, B. (2020). Single-Cell Metabolomics by Mass Spectrometry. Methods Mol Biol 2064,
8142 1-8.
- 8143 Shrestha, B. and Vertes, A. (2009). In Situ Metabolic Profiling of Single Cells by Laser Ablation
8144 Electrospray Ionization Mass Spectrometry. Anal Chem 81, 8265-8271.
- 8145 Shvartsburg, A.A., Li, F., Tang, K. and Smith, R.D. (2006). High-Resolution Field Asymmetric
8146 Waveform Ion Mobility Spectrometry Using New Planar Geometry Analyzers. Anal Chem 78,
8147 3706-3714.
- 8148 Sidore, A.M., Lan, F., Lim, S.W. and Abate, A.R. (2016). Enhanced sequencing coverage with
8149 digital droplet multiple displacement amplification. Nucleic Acids Res 44, e66.
- 8150 Simon, L.M., Wang, Y.Y. and Zhao, Z. (2021). Integration of millions of transcriptomes using
8151 batch-aware triplet neural networks. Nature Machine Intelligence 3, 705-715.
- 8152 Singer, S.J. (1992). Intercellular communication and cell-cell adhesion. Science 255, 1671-1677.
- 8153 Sinha, D., Kumar, A., Kumar, H., Bandyopadhyay, S. and Sengupta, D. (2018). dropClust:
8154 efficient clustering of ultra-large scRNA-seq data. Nucleic Acids Res 46, e36.
- 8155 Sinjab, A., Han, G., Treekitkarnmongkol, W., Hara, K., Brennan, P.M., Dang, M., Hao, D., Wang,
8156 R., Dai, E., Dejima, H., *et al.* (2021). Resolving the Spatial and Cellular Architecture of Lung
8157 Adenocarcinoma by Multiregion Single-Cell Sequencing. Cancer Discov 11, 2506-2523.
- 8158 Sinnamon, J.R., Torkenczy, K.A., Linhoff, M.W., Vitak, S.A., Mulqueen, R.M., Pliner, H.A.,
8159 Trapnell, C., Steemers, F.J., Mandel, G. and Adey, A.C. (2019). The accessible chromatin
8160 landscape of the murine hippocampus at single-cell resolution. Genome Res 29, 857-869.
- 8161 Skene, P.J. and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution
8162 mapping of DNA binding sites. Elife 6, e21856.
- 8163 Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R.,
8164 Stegle, O., Reik, W. and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for
8165 assessing epigenetic heterogeneity. Nat Methods 11, 817-820.

- 8166 Smieja, M., Wolczyk, M., Tabor, J. and Geiger, B.C. (2021). SeGMA: Semi-Supervised Gaussian
 8167 Mixture Autoencoder. IEEE Trans Neural Netw Learn Syst 32, 3930-3941.
- 8168 Smith, T., Heger, A. and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique
 8169 Molecular Identifiers to improve quantification accuracy. Genome Res 27, 491-499.
- 8170 Soneson, C. and Robinson, M.D. (2018). Bias, robustness and scalability in single-cell differential
 8171 expression analysis. Nat Methods 15, 255-261.
- 8172 Song, D. and Li, J.J. (2021). PseudotimeDE: inference of differential gene expression along cell
 8173 pseudotime with well-calibrated p-values from single-cell RNA sequencing data. Genome
 8174 Biol 22, 124.
- 8175 Song, D., Li, K., Ge, X. and Li, J.J. (2023a). ClusterDE: a post-clustering differential expression
 8176 (DE) method robust to false-positive inflation caused by double dipping. bioRxiv, doi:
 8177 10.1101/2023.07.21.550107.
- 8178 Song, D., Li, K., Hemminger, Z., Wollman, R. and Li, J.J. (2021a). scPNMF: sparse gene
 8179 encoding of single cells to facilitate gene selection for targeted gene profiling. Bioinformatics
 8180 37, i358-i366.
- 8181 Song, D., Wang, Q., Yan, G., Liu, T., Sun, T. and Li, J.J. (2023b). scDesign3 generates realistic in
 8182 silico data for multimodal single-cell and spatial omics. Nat Biotechnol 42, 247-252.
- 8183 Song, J., Liu, Y., Zhang, X., Wu, Q., Gao, J., Wang, W., Li, J., Song, Y. and Yang, C. (2021b).
 8184 Entropy subspace separation-based clustering for noise reduction (ENCORE) of scRNA-seq
 8185 data. Nucleic Acids Res 49, e18.
- 8186 Song, K., Yang, X., An, G., Xia, X., Zhao, J., Xu, X., Wan, C., Liu, T., Zheng, Y., Ren, S., *et al.*
 8187 (2022a). Targeting APLN/APJ restores blood-testis barrier and improves spermatogenesis in
 8188 murine and human diabetic models. Nat Commun 13, 7335.
- 8189 Song, L. and Crawford, G.E. (2010). DNase-seq: a high-resolution technique for mapping active
 8190 gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc
 8191 2010, pdb.prot5384.
- 8192 Song, Q., Ni, K., Liu, M., Li, Y., Wang, L., Wang, Y., Liu, Y., Yu, Z., Qi, Y., Lu, Z., *et al.* (2020).
 8193 Direct-seq: programmed gRNA scaffold for streamlined scRNA-seq in CRISPR screen.
 8194 Genome Biol 21, 136.
- 8195 Song, Q. and Su, J. (2021). DSTG: deconvoluting spatial transcriptomics data through
 8196 graph-based artificial intelligence. Brief Bioinform 22, bbaa414.
- 8197 Song, Q., Su, J. and Zhang, W. (2021c). scGCN is a graph convolutional networks algorithm for
 8198 knowledge transfer in single cell omics. Nat Commun 12, 3826.
- 8199 Song, Q., Zhu, X., Jin, L., Chen, M., Zhang, W. and Su, J. (2022b). SMGR: a joint statistical
 8200 method for integrative analysis of single-cell multi-omics data. NAR Genom Bioinform 4,
 8201 lqac056.
- 8202 Soumillon, M., Cacchiarelli, D., Semrau, S., Oudenaarden, A.v. and Mikkelsen, T.S. (2014).
 8203 Characterization of directed differentiation by high-throughput single-cell RNA-Seq. bioRxiv,
 8204 doi: 10.1101/003236, 003236.
- 8205 Specht, H., Emmott, E., Petelski, A.A., Huffman, R.G., Perlman, D.H., Serra, M., Kharchenko, P.,
 8206 Koller, A. and Slavov, N. (2021). Single-cell proteomic and transcriptomic analysis of
 8207 macrophage heterogeneity using SCoPE2. Genome Biol 22, 50.
- 8208 Specht, H., Harmange, G., Dh, P., Emmott, E., Niziolet, Z., Budnik, B. and Slavov, N. (2018).
 8209 Automated sample preparation for high-throughput single-cell proteomics. bioRxiv, doi:

- 8210 10.1101/399774.
- 8211 Srivastava, A., Sarkar, H., Gupta, N. and Patro, R. (2016). RapMap: a rapid, sensitive and accurate
8212 tool for mapping RNA-seq reads to transcriptomes. Bioinformatics 32, i192-i200.
- 8213 Stadlmann, J., Hudecz, O., Krššáková, G., Ctortecka, C., Van Raemdonck, G., Op De Beeck, J.,
8214 Desmet, G., Penninger, J.M., Jacobs, P. and Mechteder, K. (2019). Improved Sensitivity in
8215 Low-Input Proteomics Using Micropillar Array-Based Chromatography. Anal Chem 91,
8216 14203-14207.
- 8217 Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S.,
8218 Asp, M., Westholm, J.O., Huss, M., *et al.* (2016). Visualization and analysis of gene
8219 expression in tissue sections by spatial transcriptomics. Science 353, 78-82.
- 8220 Stanojevic, S., Li, Y. and Garmire, L.X. (2022). Computational methods for single-cell
8221 multi-omics integration and alignment. Genomics Proteomics Bioinformatics, 20, 836-849
- 8222 Stark, S.G., Ficek, J., Locatello, F., Bonilla, X., Chevrier, S., Singer, F., Tumor Profiler, C., Ratsch,
8223 G. and Lehmann, K.V. (2020). SCIM: universal single-cell matching with unpaired feature
8224 sets. Bioinformatics 36, i919-i927.
- 8225 Stegle, O., Teichmann, S.A. and Marioni, J.C. (2015). Computational and analytical challenges in
8226 single-cell transcriptomics. Nat Rev Genet 16, 133-145.
- 8227 Stelzer, Y., Shivalila, C.S., Soldner, F., Markoulaki, S. and Jaenisch, R. (2015). Tracing dynamic
8228 changes of DNA methylation at single-cell resolution. Cell 163, 218-229.
- 8229 Stickels, R.R., Murray, E., Kumar, P., Li, J., Marshall, J.L., Di Bella, D.J., Arlotta, P., Macosko,
8230 E.Z. and Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution
8231 with Slide-seqV2. Nat Biotechnol 39, 313-319.
- 8232 Stirparo, G.G., Boroviak, T., Guo, G., Nichols, J., Smith, A. and Bertone, P. (2018). Integrated
8233 analysis of single-cell embryo data yields a unified transcriptome signature for the human
8234 pre-implantation epiblast. Development 145, dev158501.
- 8235 Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K.,
8236 Swerdlow, H., Satija, R. and Smibert, P. (2017). Simultaneous epitope and transcriptome
8237 measurement in single cells. Nat Methods 14, 865-868.
- 8238 Stopka, S.A., Khattar, R., Agtuca, B.J., Anderton, C.R., Paša-Tolić, L., Stacey, G. and Vertes, A.
8239 (2018). Metabolic Noise and Distinct Subpopulations Observed by Single Cell LAESI Mass
8240 Spectrometry of Plant Cells *in situ*. Front Plant Sci 9, 1646.
- 8241 Storrs, E.P., Zhou, D.C., Wendl, M.C., Wyczalkowski, M.A., Karpova, A., Wang, L.B., Li, Y.,
8242 Southard-Smith, A., Jayasinghe, R.G., Yao, L., *et al.* (2022). Pollock: fishing for cell states.
8243 Bioinform Adv 2, vbac028.
- 8244 Streets, A.M., Zhang, X., Cao, C., Pang, Y., Wu, X., Xiong, L., Yang, L., Fu, Y., Zhao, L., Tang, F.,
8245 *et al.* (2014). Microfluidic single-cell whole-transcriptome sequencing. Proc Natl Acad Sci U
8246 S A 111, 7048-7053.
- 8247 Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y.,
8248 Stoeckius, M., Smibert, P. and Satija, R. (2019). Comprehensive Integration of Single-Cell
8249 Data. Cell 177, 1888-1902.e1821.
- 8250 Stuart, T., Srivastava, A., Madad, S., Lareau, C.A. and Satija, R. (2021). Single-cell chromatin
8251 state analysis with Signac. Nat Methods 18, 1333-1341.
- 8252 Su, J.H., Zheng, P., Kinrot, S.S., Bintu, B. and Zhuang, X. (2020). Genome-Scale Imaging of the
8253 3D Organization and Transcriptional Activity of Chromatin. Cell 182, 1641-1659.e1626.

- 8254 Su, K., Yu, T. and Wu, H. (2021). Accurate feature selection improves single-cell RNA-seq cell
 8255 clustering. *Brief Bioinform* 22, bbab034.
- 8256 Subramanian, A., Alperovich, M., Yang, Y. and Li, B. (2022). Biology-inspired data-driven quality
 8257 control for scientific discovery in single-cell transcriptomics. *Genome Biol* 23, 267.
- 8258 Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich,
 8259 A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.* (2005). Gene set enrichment analysis: a
 8260 knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad
 8261 Sci U S A* 102, 15545-15550.
- 8262 Sun, B. and Kumar, S. (2022). Protein Adsorption Loss—The Bottleneck of Single-Cell
 8263 Proteomics. *J Proteome Res* 21, 1808-1815.
- 8264 Sun, D., Liu, Z., Li, T., Wu, Q. and Wang, C. (2022). STRIDE: accurately decomposing and
 8265 integrating spatial transcriptomics using single-cell RNA sequencing. *Nucleic Acids Res* 50,
 8266 e42.
- 8267 Sun, S., Zhu, J. and Zhou, X. (2020a). Statistical analysis of spatial expression patterns for
 8268 spatially resolved transcriptomic studies. *Nat Methods* 17, 193-200.
- 8269 Sun, T., Song, D., Li, W.V. and Li, J.J. (2021). scDesign2: a transparent simulator that generates
 8270 high-fidelity single-cell gene expression count data with gene correlations captured. *Genome
 8271 Biol* 22, 163.
- 8272 Sun, W., Dong, H., Balaz, M., Slyper, M., Drokhlyansky, E., Colleluori, G., Giordano, A.,
 8273 Kovanicova, Z., Stefanicka, P., Balazova, L., *et al.* (2020b). snRNA-seq reveals a
 8274 subpopulation of adipocytes that regulates thermogenesis. *Nature* 587, 98-102.
- 8275 Sun, X., Liu, Y. and An, L. (2020c). Ensemble dimensionality reduction and feature gene
 8276 extraction for single-cell RNA-seq data. *Nat Commun* 11, 5853.
- 8277 Suo, C., Dann, E., Goh, I., Jardine, L., Kleshchevnikov, V., Park, J.E., Botting, R.A., Stephenson,
 8278 E., Engelbert, J., Tuong, Z.K., *et al.* (2022). Mapping the developing human immune system
 8279 across organs. *Science* 376, eabo0510.
- 8280 Svensson, V., Natarajan, K.N., Ly, L.H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A.
 8281 and Teichmann, S.A. (2017). Power analysis of single-cell RNA-sequencing experiments.
 8282 *Nat Methods* 14, 381-387.
- 8283 Svensson, V., Teichmann, S.A. and Stegle, O. (2018). SpatialDE: identification of spatially
 8284 variable genes. *Nat Methods* 15, 343-346.
- 8285 Tajik, M., Baharfar, M. and Donald, W.A. (2022). Single-cell mass spectrometry. *Trends
 8286 Biotechnol* 40, 1374-1392.
- 8287 Talwar, D., Mongia, A., Sengupta, D. and Majumdar, A. (2018). AutoImpute: Autoencoder based
 8288 imputation of single-cell RNA-seq data. *Sci Rep* 8, 16329.
- 8289 Tan, L., Xing, D., Chang, C.H., Li, H. and Xie, X.S. (2018). Three-dimensional genome structures
 8290 of single diploid human cells. *Science* 361, 924-928.
- 8291 Tanevski, J., Flores, R.O.R., Gabor, A., Schapiro, D. and Saez-Rodriguez, J. (2022). Explainable
 8292 multiview framework for dissecting spatial relationships from highly multiplexed data.
 8293 *Genome Biol* 23, 97.
- 8294 Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch,
 8295 B.B., Siddiqui, A., *et al.* (2009). mRNA-Seq whole-transcriptome analysis of a single cell.
 8296 *Nat Methods* 6, 377-382.
- 8297 Tang, L., Peng, S., Bi, Y., Shan, P. and Hu, X. (2014). A new method combining LDA and PLS for

- 8298 dimension reduction. PLoS One 9, e96944.
- 8299 Tang, W., Bertaux, F., Thomas, P., Stefanelli, C., Saint, M., Marguerat, S. and Shahrezaei, V.
8300 (2020). bayNorm: Bayesian gene expression recovery, imputation and normalization for
8301 single-cell RNA-sequencing data. Bioinformatics 36, 1174-1181.
- 8302 Tang, Z., Zhang, T., Yang, B., Su, J. and Song, Q. (2022). SiGra: Single-cell spatial elucidation
8303 through image-augmented graph transformer. bioRxiv, doi: 10.1101/2022.08.18.504464.
- 8304 Taukulis, I.A., Olszewski, R.T., Korrapati, S., Fernandez, K.A., Boger, E.T., Fitzgerald, T.S.,
8305 Morell, R.J., Cunningham, L.L. and Hoa, M. (2021). Single-Cell RNA-Seq of
8306 Cisplatin-Treated Adult Stria Vascularis Identifies Cell Type-Specific Regulatory Networks
8307 and Novel Therapeutic Gene Targets. Front Mol Neurosci 14, 718241.
- 8308 Taylor, M.J., Lukowski, J.K. and Anderton, C.R. (2021). Spatially Resolved Mass Spectrometry at
8309 the Single Cell: Recent Innovations in Proteomics and Metabolomics. J Am Soc Mass
8310 Spectrom 32, 872-894.
- 8311 Tegowski, M., Flamand, M.N. and Meyer, K.D. (2022). scDART-seq reveals distinct m(6)A
8312 signatures and mRNA methylation heterogeneity in single cells. Mol Cell 82, 868-878.e810.
- 8313 Telenius, H., Carter, N.P., Bebb, C.E., Nordenskjöld, M., Ponder, B.A. and Tunnacliffe, A. (1992).
8314 Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single
8315 degenerate primer. Genomics 13, 718-725.
- 8316 Teng, H., Yuan, Y. and Bar-Joseph, Z. (2022). Clustering spatial transcriptomics data.
8317 Bioinformatics 38, 997-1004.
- 8318 Thorner, K., Zorn, A.M. and Chaturvedi, P. (2021). ELeFHAnt: A supervised machine learning
8319 approach for label harmonization and annotation of single cell RNA-seq data. bioRxiv, doi:
8320 10.1101/2021.09.07.459342.
- 8321 Tian, Y., Li, Q., Yang, Z., Zhang, S., Xu, J., Wang, Z., Bai, H., Duan, J., Zheng, B., Li, W., et al.
8322 (2022). Single-cell transcriptomic profiling reveals the tumor heterogeneity of small-cell lung
8323 cancer. Signal Transduct Target Ther 7, 346.
- 8324 Tiberi, S., Crowell, H.L., Samartsidis, P., Weber, L.M. and Robinson, M.D. (2022). *distinct*: a
8325 novel approach to differential distribution analyses. bioRxiv, doi:
8326 10.1101/2020.11.24.394213.
- 8327 Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A.,
8328 Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of
8329 metastatic melanoma by single-cell RNA-seq. Science 352, 189-196.
- 8330 Tobler, K.J., Brezina, P.R., Benner, A.T., Du, L., Xu, X. and Kearns, W.G. (2014). Two different
8331 microarray technologies for preimplantation genetic diagnosis and screening, due to
8332 reciprocal translocation imbalances, demonstrate equivalent euploidy and clinical pregnancy
8333 rates. J Assist Reprod Genet 31, 843-850.
- 8334 Torres, A.J., Hill, A.S. and Love, J.C. (2014). Nanowell-based immunoassays for measuring
8335 single-cell secretion: characterization of transport and surface binding. Anal Chem 86,
8336 11562-11569.
- 8337 Townes, F.W., Hicks, S.C., Aryee, M.J. and Irizarry, R.A. (2019). Feature selection and dimension
8338 reduction for single-cell RNA-Seq based on a multinomial model. Genome Biol 20, 295.
- 8339 Tracy, S., Yuan, G.C. and Dries, R. (2019). RESCUE: imputing dropout events in single-cell
8340 RNA-sequencing data. BMC Bioinformatics 20, 388.
- 8341 Tran, H.T.N., Ang, K.S., Chevrier, M., Zhang, X., Lee, N.Y.S., Goh, M. and Chen, J. (2020). A

- benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 21, 12.
- Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S. and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 32, 381-386.
- Trapnell, C., Pachter, L. and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105-1111.
- Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28, 511-515.
- Treff, N.R., Fedick, A., Tao, X., Devkota, B., Taylor, D. and Scott, R.T., Jr. (2013). Evaluation of targeted next-generation sequencing-based preimplantation genetic diagnosis of monogenic disease. *Fertil Steril* 99, 1377-1384.e1376.
- Treff, N.R., Tao, X., Ferry, K.M., Su, J., Taylor, D. and Scott, R.T., Jr. (2012). Development and validation of an accurate quantitative real-time polymerase chain reaction-based assay for human blastocyst comprehensive chromosomal aneuploidy screening. *Fertil Steril* 97, 819-824.
- Tsai, C.-F., Zhang, P., Scholten, D., Martin, K., Wang, Y.-T., Zhao, R., Chrisler, W.B., Patel, D.B., Dou, M., Jia, Y., et al. (2021). Surfactant-assisted one-pot sample preparation for label-free single-cell proteomics. *Commun Biol* 4, 265.
- Tsai, C.-F., Zhao, R., Williams, S.M., Moore, R.J., Schultz, K., Chrisler, W.B., Pasa-Tolic, L., Rodland, K.D., Smith, R.D., Shi, T., et al. (2020). An Improved Boosting to Amplify Signal with Isobaric Labeling (iBASIL) Strategy for Precise Quantitative Single-cell Proteomics. *Mol Cell Proteomics* 19, 828-838.
- Tu, Q., Cameron, R.A., Worley, K.C., Gibbs, R.A. and Davidson, E.H. (2012). Gene structure in the sea urchin *Strongylocentrotus purpuratus* based on transcriptome analysis. *Genome Res* 22, 2079-2087.
- Tung, P.Y., Blischak, J.D., Hsiao, C.J., Knowles, D.A., Burnett, J.E., Pritchard, J.K. and Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. *Sci Rep* 7, 39921.
- Ulirsch, J.C., Lareau, C.A., Bao, E.L., Ludwig, L.S., Guo, M.H., Benner, C., Satpathy, A.T., Kartha, V.K., Salem, R.M., Hirschhorn, J.N., et al. (2019). Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat Genet* 51, 683-693.
- Unger, M.A., Chou, H.P., Thorsen, T., Scherer, A. and Quake, S.R. (2000). Monolithic microfabricated valves and pumps by multilayer soft lithography. *Science* 288, 113-116.
- Uzun, Y., Wu, H. and Tan, K. (2021). Predictive modeling of single-cell DNA methylome data enhances integration with transcriptome data. *Genome Res* 31, 101-109.
- Vaisvila, R., Ponnaluri, V.K.C., Sun, Z., Langhorst, B.W., Saleh, L., Guan, S., Dai, N., Campbell, M.A., Sexton, B.S., Marks, K., et al. (2021). Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res* 31, 1280-1289.
- Vallejos, C.A., Marioni, J.C. and Richardson, S. (2015). BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Comput Biol* 11, e1004333.

- van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., *et al.* (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* 174, 716-729.e727.
- Vandenbon, A. and Diez, D. (2020). A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. *Nat Commun* 11, 4318.
- Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråle, L., Navarro, J.F., *et al.* (2019). High-definition spatial transcriptomics for *in situ* tissue profiling. *Nat Methods* 16, 987-990.
- Villas-Bôas, S.G., Mas, S., Akesson, M., Smedsgaard, J. and Nielsen, J. (2005). Mass spectrometry in metabolome analysis. *Mass Spectrom Rev* 24, 613-646.
- Vitak, S.A., Torkenczy, K.A., Rosenkrantz, J.L., Fields, A.J., Christiansen, L., Wong, M.H., Carbone, L., Steemers, F.J. and Adey, A. (2017). Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat Methods* 14, 302-308.
- Wagner, F. (2020). Monet: An open-source Python package for analyzing and integrating scRNA-Seq data using PCA-based latent spaces. *bioRxiv*, doi: 10.1101/2020.06.08.140673.
- Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. and Batzoglou, S. (2017a). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Methods* 14, 414-416.
- Wang, C., Sun, D., Huang, X., Wan, C., Li, Z., Han, Y., Qin, Q., Fan, J., Qiu, X., Xie, Y., *et al.* (2020). Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol* 21, 198.
- Wang, C.X., Zhang, L. and Wang, B. (2022a). One Cell At a Time (OCAT): a unified framework to integrate and analyze single-cell RNA-seq data. *Genome Biol* 23, 102.
- Wang, D., Hou, S., Zhang, L., Wang, X., Liu, B. and Zhang, Z. (2021a). iMAP: integration of multiple single-cell datasets by adversarial paired transfer networks. *Genome Biol* 22, 63.
- Wang, J., Agarwal, D., Huang, M., Hu, G., Zhou, Z., Ye, C. and Zhang, N.R. (2019a). Data denoising with transfer learning in single-cell transcriptomics. *Nat Methods* 16, 875-878.
- Wang, J., Fan, H.C., Behr, B. and Quake, S.R. (2012). Genome-wide single-cell analysis of recombination activity and *de novo* mutation rates in human sperm. *Cell* 150, 402-412.
- Wang, J., Qian, J., Hoeksema, M.D., Zou, Y., Espinosa, A.V., Rahman, S.M., Zhang, B. and Massion, P.P. (2013). Integrative genomics analysis identifies candidate drivers at 3q26-29 amplicon in squamous cell carcinoma of the lung. *Clin Cancer Res* 19, 5580-5590.
- Wang, J., Vasaikar, S., Shi, Z., Greer, M. and Zhang, B. (2017b). WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res* 45, W130-w137.
- Wang, K., Kumar, T., Wang, J., Minussi, D.C., Sei, E., Li, J., Tran, T.M., Thennavan, A., Hu, M., Casasent, A.K., *et al.* (2023a). Archival single-cell genomics reveals persistent subclones during DCIS progression. *Cell* 186, 3968-3982.e3915.
- Wang, K., Li, X., Dong, S., Liang, J., Mao, F., Zeng, C., Wu, H., Wu, J., Cai, W. and Sun, Z.S. (2015). Q-RRBS: a quantitative reduced representation bisulfite sequencing method for single-cell methylome analyses. *Epigenetics* 10, 775-783.
- Wang, L. (2021). Single-cell normalization and association testing unifying CRISPR screen and gene co-expression analyses with Normalisr. *Nat Commun* 12, 6395.
- Wang, L., Ma, H., Wen, Z., Niu, L., Chen, X., Liu, H., Zhang, S., Xu, J., Zhu, Y., Li, H., *et al.*

- 8430 (2023b). Single-cell RNA-sequencing reveals heterogeneity and intercellular crosstalk in
8431 human tuberculosis lung. *J Infect* 87, 373-384.
- 8432 Wang, Q., Xiong, H., Ai, S., Yu, X., Liu, Y., Zhang, J. and He, A. (2019b). CoBATCH for
8433 High-Throughput Single-Cell Epigenomic Profiling. *Mol Cell* 76, 206-216.e207.
- 8434 Wang, R., Zhao, H., Zhang, X., Zhao, X., Song, Z. and Ouyang, J. (2019c). Metabolic
8435 Discrimination of Breast Cancer Subtypes at the Single-Cell Level by Multiple
8436 Microextraction Coupled with Mass Spectrometry. *Anal Chem* 91, 3667-3674.
- 8437 Wang, S., Karikomi, M., MacLean, A.L. and Nie, Q. (2019d). Cell lineage and communication
8438 network inference via optimization for single-cell transcriptomics. *Nucleic Acids Res* 47,
8439 e66.
- 8440 Wang, S., Sun, S.T., Zhang, X.Y., Ding, H.R., Yuan, Y., He, J.J., Wang, M.S., Yang, B. and Li, Y.B.
8441 (2023c). The Evolution of Single-Cell RNA Sequencing Technology and Application:
8442 Progress and Perspectives. *Int J Mol Sci* 24, 2943.
- 8443 Wang, T., Johnson, T.S., Shao, W., Lu, Z., Helm, B.R., Zhang, J. and Huang, K. (2019e).
8444 BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch
8445 correction reveals hidden high-resolution cellular subtypes. *Genome Biol* 20, 165.
- 8446 Wang, T., Wei, J.J., Sabatini, D.M. and Lander, E.S. (2014a). Genetic screens in human cells using
8447 the CRISPR-Cas9 system. *Science* 343, 80-84.
- 8448 Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N., Vesuna, S., Evans, K., Liu,
8449 C., Ramakrishnan, C., Liu, J., *et al.* (2018). Three-dimensional intact-tissue sequencing of
8450 single-cell transcriptional states. *Science* 361, eaat5691.
- 8451 Wang, Y., Chen, L., Jo, J. and Wang, Y. (2022b). Joint t-SNE for Comparable Projections of
8452 Multiple High-Dimensional Datasets. *IEEE Trans Vis Comput Graph* 28, 623-632.
- 8453 Wang, Y., Gao, J., Xuan, C., Guan, T., Wang, Y., Zhou, G. and Ding, T. (2022c). FSCAM:
8454 CAM-Based Feature Selection for Clustering scRNA-seq. *Interdiscip Sci* 14, 394-408.
- 8455 Wang, Y., Guan, Z.-Y., Shi, S.-W., Jiang, Y.-R., Wu, Q., Wu, J., Chen, J.-B., Ying, W.-X., Xu,
8456 Q.-Q., Fan, Q.-X., *et al.* (2022d). Pick-up Single-Cell Proteomic Analysis for Quantifying up
8457 to 3000 Proteins in a Tumor Cell. *bioRxiv*, doi: 10.1101/2022.06.28.498038.
- 8458 Wang, Y., Liu, T. and Zhao, H. (2022e). ResPAN: a powerful batch correction model for
8459 scRNA-seq data through residual adversarial networks. *Bioinformatics* 38, 3942-3949.
- 8460 Wang, Y., Song, B., Wang, S., Chen, M., Xie, Y., Xiao, G., Wang, L. and Wang, T. (2022f). Sprod
8461 for de-noising spatially resolved transcriptomics data based on position and image
8462 information. *Nat Methods* 19, 950-958.
- 8463 Wang, Y., Waters, J., Leung, M.L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S.,
8464 Liang, H., *et al.* (2014b). Clonal evolution in breast cancer revealed by single nucleus
8465 genome sequencing. *Nature* 512, 155-160.
- 8466 Wang, Y., Xie, S., Armendariz, D. and Hon, G.C. (2022g). Computational identification of clonal
8467 cells in single-cell CRISPR screens. *BMC Genomics* 23, 135.
- 8468 Wang, Y., Yuan, P., Yan, Z., Yang, M., Huo, Y., Nie, Y., Zhu, X., Qiao, J. and Yan, L. (2021b).
8469 Single-cell multiomics sequencing reveals the functional regulatory landscape of early
8470 embryos. *Nat Commun* 12, 1247.
- 8471 Wangwu, J., Sun, Z. and Lin, Z. (2021). scAMACE: model-based approach to the joint analysis of
8472 single-cell data on chromatin accessibility, gene expression and methylation. *Bioinformatics*
8473 37, 3874-3880.

- Wei, H., Han, T., Li, T., Wu, Q. and Wang, C. (2023). SCREE: a comprehensive pipeline for single-cell multi-modal CRISPR screen data processing and analysis. *Brief Bioinform* 24, bbad123.
- Wei, X., Li, Z., Ji, H. and Wu, H. (2022). EDClust: an EM-MM hybrid method for cell clustering in multiple-subject single-cell RNA sequencing. *Bioinformatics* 38, 2692-2699.
- Wei, Z., Zhang, X., Si, X., Gong, X., Zhang, S. and Zhang, X. (2020). Development of Pico-ESI-MS for Single-Cell Metabolomics Analysis. *Methods Mol Biol* 2064, 31-59.
- Weibel, K.E., Mor, J.R. and Fiechter, A. (1974). Rapid sampling of yeast cells and automated assays of adenylate, citrate, pyruvate and glucose-6-phosphate pools. *Anal Biochem* 58, 208-216.
- Welch, J.D., Hartemink, A.J. and Prins, J.F. (2017). MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol* 18, 138.
- Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C. and Macosko, E.Z. (2019). Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* 177, 1873-1887.e1817.
- Wells, D., Escudero, T., Levy, B., Hirschhorn, K., Delhanty, J.D. and Munné, S. (2002). First clinical application of comparative genomic hybridization and polar body testing for preimplantation genetic diagnosis of aneuploidy. *Fertil Steril* 78, 543-549.
- Wells, D., Kaur, K., Grifo, J., Glassner, M., Taylor, J.C., Fragouli, E. and Munne, S. (2014). Clinical utilisation of a rapid low-pass whole genome sequencing technique for the diagnosis of aneuploidy in human embryos prior to implantation. *J Med Genet* 51, 553-562.
- Wen, Y., Zhao, J., Zhang, R., Liu, F., Chen, X., Wu, D., Wang, M., Liu, C., Su, P., Meng, P., *et al.* (2024). Identification and characterization of human hematopoietic mesoderm. *Sci China Life Sci* 67, 320-331.
- Wen, Z.H., Langsam, J.L., Zhang, L., Shen, W. and Zhou, X. (2022). A Bayesian factorization method to recover single-cell RNA sequencing data. *Cell Rep Methods* 2, 100133.
- Wickham, H. (2011). ggplot2: Elegant Graphics for Data Analysis. *Journal of The Royal Statistical Society Series A-statistics in Society* 174, 245-246.
- Wilk, A.J., Shalek, A.K., Holmes, S. and Blish, C.A. (2023). Comparative analysis of cell-cell communication at single-cell resolution. *bioRxiv*, doi: 10.1101/s41587-023-01782-z.
- Williams, S.M., Liyu, A.V., Tsai, C.-F., Moore, R.J., Orton, D.J., Chrisler, W.B., Gaffrey, M.J., Liu, T., Smith, R.D., Kelly, R.T., *et al.* (2020). Automated Coupling of Nanodroplet Sample Preparation with Liquid Chromatography-Mass Spectrometry for High-Throughput Single-Cell Proteomics. *Anal Chem* 92, 10588-10596.
- Wilson, P.C., Wu, H., Kirita, Y., Uchimura, K., Ledru, N., Rennke, H.G., Welling, P.A., Waikar, S.S. and Humphreys, B.D. (2019). The single-cell transcriptomic landscape of early human diabetic nephropathy. *Proc Natl Acad Sci U S A* 116, 19619-19625.
- Wilson, S.B., Howden, S.E., Vanslambrouck, J.M., Dorison, A., Alquicira-Hernandez, J., Powell, J.E. and Little, M.H. (2022). DevKidCC allows for robust classification and direct comparisons of kidney organoid datasets. *Genome Med* 14, 19.
- Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., *et al.* (2007). HMDB: the Human Metabolome Database. *Nucleic Acids Res* 35, D521-526.

- 8518 Wiśniewski, J.R., Hein, M.Y., Cox, J. and Mann, M. (2014). A “Proteomic Ruler” for Protein
 8519 Copy Number and Concentration Estimation without Spike-in Standards. Mol Cell
 8520 Proteomics 13, 3497-3506.
- 8521 Wolf, F.A., Angerer, P. and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression
 8522 data analysis. Genome Biol 19, 15.
- 8523 Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Götgens, B., Rajewsky, N., Simon, L.
 8524 and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory
 8525 inference through a topology preserving map of single cells. Genome Biol 20, 59.
- 8526 Wolock, S.L., Lopez, R. and Klein, A.M. (2019). Scrublet: Computational Identification of Cell
 8527 Doublets in Single-Cell Transcriptomic Data. Cell Syst 8, 281-291.e289.
- 8528 Woo, J., Williams, S.M., Markillie, L.M., Feng, S., Tsai, C.F., Aguilera-Vazquez, V., Sontag, R.L.,
 8529 Moore, R.J., Hu, D., Mehta, H.S., *et al.* (2021). High-throughput and high-efficiency sample
 8530 preparation for single-cell proteomics using a nested nanowell chip. Nat Commun 12, 6246.
- 8531 Wroblewska, A., Dhainaut, M., Ben-Zvi, B., Rose, S.A., Park, E.S., Amir, E.D., Bektesevic, A.,
 8532 Baccarini, A., Merad, M., Rahman, A.H., *et al.* (2018). Protein Barcodes Enable
 8533 High-Dimensional Single-Cell CRISPR Screens. Cell 175, 1141-1155.e1116.
- 8534 Wu B, Wang Y, Yan J, Liu M, Li X, Tang F, Bao S. (2024) Blastoids generated purely from
 8535 embryonic stem cells both in mice and humans. Sci. China Life Sci., 67, 418 - 420.
- 8536 Wu, H., Wu, Y., Jiang, Y., Zhou, B., Zhou, H., Chen, Z., Xiong, Y., Liu, Q. and Zhang, H. (2022).
 8537 scHiCStackL: a stacking ensemble learning-based method for single-cell Hi-C classification
 8538 using cell embedding. Brief Bioinform 23, bbab396.
- 8539 Wu, K., Jia, F., Zheng, W., Luo, Q., Zhao, Y. and Wang, F. (2017a). Visualization of metallodrugs
 8540 in single cells by secondary ion mass spectrometry imaging. J Biol Inorg Chem 22, 653-661.
- 8541 Wu, L., Yan, J., Bai, Y., Chen, F., Xu, J., Zou, X., Huang, A., Hou, L., Zhong, Y., Jing, Z., *et al.*
 8542 (2021a). Spatially-resolved transcriptomics analyses of invasive fronts in solid tumors.
 8543 bioRxiv, doi: 10.1101/2021.10.21.465135.
- 8544 Wu, P., Gao, Y., Guo, W. and Zhu, P. (2019a). Using local alignment to enhance single-cell
 8545 bisulfite sequencing data efficiency. Bioinformatics 35, 3273-3278.
- 8546 Wu, P.H., Gilkes, D.M., Phillip, J.M., Narkar, A., Cheng, T.W., Marchand, J., Lee, M.H., Li, R.
 8547 and Wirtz, D. (2020). Single-cell morphology encodes metastatic potential. Sci Adv 6,
 8548 eaaw6938.
- 8549 Wu, R., Guo, W., Qiu, X., Wang, S., Sui, C., Lian, Q., Wu, J., Shan, Y., Yang, Z., Yang, S., *et al.*
 8550 (2021b). Comprehensive analysis of spatial architecture in primary liver cancer. Sci Adv 7,
 8551 eabg3750.
- 8552 Wu, R., Xing, S., Badv, M., Didar, T.F. and Lu, Y. (2019b). Step-Wise Assessment and
 8553 Optimization of Sample Handling Recovery Yield for Nanoproteomic Analysis of 1000
 8554 Mammalian Cells. Anal Chem 91, 10395-10400.
- 8555 Wu, S.J., Furlan, S.N., Mihalas, A.B., Kaya-Okur, H.S., Feroze, A.H., Emerson, S.N., Zheng, Y.,
 8556 Carson, K., Cimino, P.J., Keene, C.D., *et al.* (2021c). Single-cell CUT&Tag analysis of
 8557 chromatin modifications in differentiation and tumor progression. Nat Biotechnol 39,
 8558 819-824.
- 8559 Wu, S.Z., Al-Eryani, G., Roden, D.L., Junankar, S., Harvey, K., Andersson, A., Thennavan, A.,
 8560 Wang, C., Torpy, J.R., Bartonicek, N., *et al.* (2021d). A single-cell and spatially resolved atlas
 8561 of human breast cancers. Nat Genet 53, 1334-1347.

- 8562 Wu, W., Liu, Y., Dai, Q., Yan, X. and Wang, Z. (2021e). G2S3: A gene graph-based imputation
8563 method for single-cell RNA sequencing data. PLoS Comput Biol 17, e1009029.
- 8564 Wu, X., Inoue, A., Suzuki, T. and Zhang, Y. (2017b). Simultaneous mapping of active DNA
8565 demethylation and sister chromatid exchange in single cells. Genes Dev 31, 511-523.
- 8566 Xi, Y. and Li, W. (2009). BSMP: whole genome bisulfite sequence MAPping program. BMC
8567 Bioinformatics 10, 232.
- 8568 Xia, C., Fan, J., Emanuel, G., Hao, J. and Zhuang, X. (2019). Spatial transcriptome profiling by
8569 MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene
8570 expression. Proc Natl Acad Sci U S A 116, 19490-19499.
- 8571 Xie, H. and Ding, X. (2022). The Intriguing Landscape of Single-Cell Protein Analysis. Adv Sci
8572 (Weinh) 9, e2105932.
- 8573 Xie, Q., Han, C., Jin, V. and Lin, S. (2022). HiCImpute: A Bayesian hierarchical model for
8574 identifying structural zeros and enhancing single cell Hi-C data. PLoS Comput Biol 18,
8575 e1010129.
- 8576 Xie, S., Duan, J., Li, B., Zhou, P. and Hon, G.C. (2017). Multiplexed Engineering and Analysis of
8577 Combinatorial Enhancer Activity in Single Cells. Mol Cell 66, 285-299.e285.
- 8578 Xin, Y., Lyu, P., Jiang, J., Zhou, F., Wang, J., Blackshaw, S. and Qian, J. (2022). LRLoop:
8579 Feedback loops as a design principle of cell-cell communication. bioRxiv, doi:
8580 10.1101/2022.02.04.479174.
- 8581 Xing, D., Tan, L., Chang, C.H., Li, H. and Xie, X.S. (2021). Accurate SNV detection in single
8582 cells by transposon-based whole-genome amplification of complementary strands. Proc Natl
8583 Acad Sci U S A 118, e2013106118.
- 8584 Xiong, K.-X., Zhou, H.-L., Yin, J.-H., Kristiansen, K., Yang, H.-M. and Li, G.-B. (2021a). Chord:
8585 Identifying Doublets in Single-Cell RNA Sequencing Data by an Ensemble Machine
8586 Learning Algorithm. bioRxiv, doi: 10.1101/2021.05.07.442884.
- 8587 Xiong, L., Tian, K., Li, Y. and Zhang, Q.C. (2021b). Online single-cell data integration through
8588 projecting heterogeneous datasets into a common cell-embedding space. bioRxiv, doi:
8589 10.1101/2021.04.06.438536.
- 8590 Xu, C. and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel
8591 clustering method. Bioinformatics 31, 1974-1980.
- 8592 Xu, J., Cai, L., Liao, B., Zhu, W. and Yang, J. (2020a). CMF-Impute: an accurate imputation tool
8593 for single-cell RNA-seq data. Bioinformatics 36, 3139-3147.
- 8594 Xu, S., Liu, M., Bai, Y. and Liu, H. (2021a). Multi-Dimensional Organic Mass Cytometry:
8595 Simultaneous Analysis of Proteins and Metabolites on Single Cells. Angew Chem Int Ed
8596 Engl 60, 1806-1812.
- 8597 Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., *et al.*
8598 (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a
8599 kidney tumor. Cell 148, 886-895.
- 8600 Xu, Y., Begoli, E. and McCord, R.P. (2021b). sciCAN: Single-cell chromatin accessibility and
8601 gene expression data integration via Cycle-consistent Adversarial Network. bioRxiv, doi:
8602 10.1101/2021.11.30.470677.
- 8603 Xu, Y. and McCord, R.P. (2022). Diagonal integration of multimodal single-cell data: potential
8604 pitfalls and paths forward. Nat Commun 13, 3505.
- 8605 Xu, Y., Zhang, Z., You, L., Liu, J., Fan, Z. and Zhou, X. (2020b). scIGANs: single-cell RNA-seq

- 8606 imputation using generative adversarial networks. Nucleic Acids Res 48, e85.
- 8607 Xu, Z., Chen, D., Hu, Y., Jiang, K., Huang, H., Du, Y., Wu, W., Wang, J., Sui, J., Wang, W., *et al.*
 8608 (2022). Anatomically distinct fibroblast subsets determine skin autoimmune patterns. Nature
 8609 601, 118-124.
- 8610 Yan, F., Zhao, Z. and Simon, L.M. (2021). EmptyNN: A neural network based on positive and
 8611 unlabeled learning to remove cell-free droplets and recover lost cells in scRNA-seq data.
 8612 Patterns (N Y) 2, 100311.
- 8613 Yang, L., Liu, J., Lu, Q., Riggs, A.D. and Wu, X. (2017). SAIC: an iterative clustering approach
 8614 for analysis of single cell RNA-seq data. BMC Genomics 18, 689.
- 8615 Yang, L., Wang, Z., Deng, Y., Li, Y., Wei, W. and Shi, Q. (2016). Single-Cell, Multiplexed Protein
 8616 Detection of Rare Tumor Cells Based on a Beads-on-Barcode Antibody Microarray. Anal
 8617 Chem 88, 11077-11083.
- 8618 Yang, L., Zhu, Y., Yu, H., Cheng, X., Chen, S., Chu, Y., Huang, H., Zhang, J. and Li, W. (2020).
 8619 scMAGECK links genotypes with multiple phenotypes in single-cell CRISPR screens.
 8620 Genome Biol 21, 19.
- 8621 Yang, S., Chen, D., Xie, L., Zou, X., Xiao, Y., Rao, L., Yao, T., Zhang, Q., Cai, L., Huang, F., *et al.*
 8622 (2023a). Developmental dynamics of the single nucleus regulatory landscape of pig
 8623 hippocampus. Sci China Life Sci 66, 2614-2628.
- 8624 Yang, X., Yan, J., Cheng, Y. and Zhang, Y. (2023b). Learning Deep Generative Clustering via
 8625 Mutual Information Maximization. IEEE Trans Neural Netw Learn Syst 34, 6263-6275.
- 8626 Yang, Y., Li, G., Qian, H., Wilhelmsen, K.C., Shen, Y. and Li, Y. (2021a). SMNN: batch effect
 8627 correction for single-cell RNA-seq data via supervised mutual nearest neighbor detection.
 8628 Brief Bioinform 22, bbaa097.
- 8629 Yang, Y., Li, G., Xie, Y., Wang, L., Lagler, T.M., Yang, Y., Liu, J., Qian, L. and Li, Y. (2021b).
 8630 iSMNN: batch effect correction for single-cell RNA-seq data via iterative supervised mutual
 8631 nearest neighbor refinement. Brief Bioinform 22, bbab122.
- 8632 Yang, Y., Shi, X., Liu, W., Zhou, Q., Chan Lau, M., Chun Tatt Lim, J., Sun, L., Ng, C.C.Y., Yeong,
 8633 J. and Liu, J. (2022). SC-MEB: spatial clustering with hidden Markov random field using
 8634 empirical Bayes. Brief Bioinform 23, bbab466.
- 8635 Yao, H., Zhao, H., Zhao, X., Pan, X., Feng, J., Xu, F., Zhang, S. and Zhang, X. (2019). Label-free
 8636 Mass Cytometry for Unveiling Cellular Metabolic Heterogeneity. Anal Chem 91, 9777-9783.
- 8637 Yao, Y.X., La, Y.F., Di, R., Liu, Q.Y., Hu, W.P., Wang, X.Y. and Chu, M.X. (2018). [Comparison
 8638 of different single cell whole genome amplification methods and MALBAC applications in
 8639 assisted reproduction]. Yi Chuan 40, 620-631.
- 8640 Yasen, A., Aini, A., Wang, H., Li, W., Zhang, C., Ran, B., Tuxun, T., Maimaitinijiati, Y., Shao, Y.,
 8641 Aji, T., *et al.* (2020). Progress and applications of single-cell sequencing techniques. Infect
 8642 Genet Evol 80, 104198.
- 8643 Yin, L., Zhang, Z., Liu, Y., Gao, Y. and Gu, J. (2019a). Recent advances in single-cell analysis by
 8644 mass spectrometry. Analyst 144, 824-845.
- 8645 Yin, R., Burnum-Johnson, K.E., Sun, X., Dey, S.K. and Laskin, J. (2019b). High spatial resolution
 8646 imaging of biological tissues using nanospray desorption electrospray ionization mass
 8647 spectrometry. Nat Protoc 14, 3445-3470.
- 8648 Yin, R., Prabhakaran, V. and Laskin, J. (2018). Quantitative Extraction and Mass Spectrometry
 8649 Analysis at a Single-Cell Level. Anal Chem 90, 7937-7945.

- 8650 Yin, Y., Jiang, Y., Lam, K.G., Berleth, J.B., Disteche, C.M., Noble, W.S., Steemers, F.J.,
 8651 Camerini-Otero, R.D., Adey, A.C. and Shendure, J. (2019c). High-Throughput Single-Cell
 8652 Sequencing with Linear Amplification. *Mol Cell* 76, 676-690.e610.
- 8653 You, Y., Tian, L., Su, S., Dong, X., Jabbari, J.S., Hickey, P.F. and Ritchie, M.E. (2021).
 8654 Benchmarking UMI-based single-cell RNA-seq preprocessing workflows. *Genome Biol* 22,
 8655 339.
- 8656 Yu, L., Cao, Y., Yang, J.Y.H. and Yang, P. (2022). Benchmarking clustering algorithms on
 8657 estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biol* 23,
 8658 49.
- 8659 Yu, M., Abnousi, A., Zhang, Y., Li, G., Lee, L., Chen, Z., Fang, R., Lagler, T.M., Yang, Y., Wen, J.,
 8660 *et al.* (2021). SnapHiC: a computational pipeline to identify chromatin loops from single-cell
 8661 Hi-C data. *Nat Methods* 18, 1056-1059.
- 8662 Yuan, H. and Kelley, D.R. (2022). scBasset: sequence-based modeling of single-cell ATAC-seq
 8663 using convolutional neural networks. *Nat Methods* 19, 1088-1096.
- 8664 Yuan, P., Xu, C., He, N., Lu, X., Zhang, X., Shang, J., Zhu, H., Gong, C., Kuang, H., Tang, T., *et*
 8665 *al.* (2023). Watermelon domestication was shaped by stepwise selection and regulation of the
 8666 metabolome. *Sci China Life Sci* 66, 579-594.
- 8667 Yuan, Y. and Bar-Joseph, Z. (2020). GCNG: graph convolutional networks for inferring gene
 8668 interaction from spatial transcriptomics data. *Genome Biol* 21, 300.
- 8669 Zahn, H., Steif, A., Laks, E., Eirew, P., VanInsberghe, M., Shah, S.P., Aparicio, S. and Hansen, C.L.
 8670 (2017). Scalable whole-genome single-cell library preparation without preamplification. *Nat
 8671 Methods* 14, 167-173.
- 8672 Zamanighomi, M., Lin, Z., Daley, T., Chen, X., Duren, Z., Schep, A., Greenleaf, W.J. and Wong,
 8673 W.H. (2018). Unsupervised clustering and epigenetic classification of single cells. *Nat
 8674 Commun* 9, 2410.
- 8675 Zappia, L., Phipson, B. and Oshlack, A. (2017). Splatter: simulation of single-cell RNA
 8676 sequencing data. *Genome Biol* 18, 174.
- 8677 Zeira, R., Land, M., Strzalkowski, A. and Raphael, B.J. (2022). Alignment and integration of
 8678 spatial transcriptomics data. *Nat Methods* 19, 567-575.
- 8679 Zelig, A. and Kaplan, N. (2020). KMD clustering: Robust generic clustering of biological data.
 8680 *bioRxiv*, doi: 10.1101/2020.10.04.325233.
- 8681 Zeng, P. and Lin, Z. (2021). coupleCoC+: An information-theoretic co-clustering-based transfer
 8682 learning framework for the integrative analysis of single-cell genomic data. *PLoS Comput
 8683 Biol* 17, e1009064.
- 8684 Zeng, P., Wangwu, J. and Lin, Z. (2021). Coupled co-clustering-based unsupervised transfer
 8685 learning for the integrative analysis of single-cell genomic data. *Brief Bioinform* 22,
 8686 bbaa347.
- 8687 Zeng, Y., Wei, Z., Pan, Z., Lu, Y. and Yang, Y. (2022a). A robust and scalable graph neural network
 8688 for accurate single-cell classification. *Brief Bioinform* 23, bbab570.
- 8689 Zeng, Y., Wei, Z., Yu, W., Yin, R., Yuan, Y., Li, B., Tang, Z., Lu, Y. and Yang, Y. (2022b). Spatial
 8690 transcriptomics prediction from histology jointly through Transformer and graph neural
 8691 networks. *Brief Bioinform* 23, bbac297.
- 8692 Zeng, Y., Wei, Z., Zhong, F., Pan, Z., Lu, Y. and Yang, Y. (2022c). A parameter-free deep
 8693 embedded clustering method for single-cell RNA-seq data. *Brief Bioinform* 23, bbac172.

- 8694 Zenobi, R. (2013). Single-cell metabolomics: analytical and biological perspectives. *Science* 342,
 8695 1243259.
- 8696 Zhang, A.W., O'Flanagan, C., Chavez, E.A., Lim, J.L.P., Ceglia, N., McPherson, A., Wiens, M.,
 8697 Walters, P., Chan, T., Hewitson, B., *et al.* (2019a). Probabilistic cell-type assignment of
 8698 single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods* 16, 1007-1015.
- 8699 Zhang, B., Kirov, S. and Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene
 8700 sets in various biological contexts. *Nucleic Acids Res* 33, W741-748.
- 8701 Zhang, B., Srivastava, A., Mimitou, E., Stuart, T., Raimondi, I., Hao, Y., Smibert, P. and Satija, R.
 8702 (2022a). Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro. *Nat
 8703 Biotechnol* 40, 1220-1230.
- 8704 Zhang, F., Wu, Y. and Tian, W. (2019b). A novel approach to remove the batch effect of single-cell
 8705 data. *Cell Discov* 5, 46.
- 8706 Zhang, K., Feng, W. and Wang, P. (2022b). Identification of spatially variable genes with graph
 8707 cuts. *Nat Commun* 13, 5488.
- 8708 Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W. and Arnheim, N. (1992). Whole genome
 8709 amplification from a single cell: implications for genetic analysis. *Proc Natl Acad Sci U S A*
 8710 89, 5847-5851.
- 8711 Zhang, L. and Nie, Q. (2021). scMC learns biological variation through the alignment of multiple
 8712 single-cell genomics datasets. *Genome Biol* 22, 10.
- 8713 Zhang, L., Sevinsky, C.J., Davis, B.M. and Vertes, A. (2018). Single-Cell Mass Spectrometry of
 8714 Subpopulations Selected by Fluorescence Microscopy. *Anal Chem* 90, 4626-4634.
- 8715 Zhang, L. and Vertes, A. (2015). Energy Charge, Redox State, and Metabolite Turnover in Single
 8716 Human Hepatocytes Revealed by Capillary Microsampling Mass Spectrometry. *Anal Chem*
 8717 87, 10397-10405.
- 8718 Zhang, L. and Vertes, A. (2018). Single-Cell Mass Spectrometry Approaches to Explore Cellular
 8719 Heterogeneity. *Angew Chem Int Ed Engl* 57, 4466-4477.
- 8720 Zhang, L., Zhang, J. and Nie, Q. (2022c). DIRECT-NET: An efficient method to discover
 8721 cis-regulatory elements and construct regulatory networks from single-cell multiomics data.
 8722 *Sci Adv* 8, eabl7393.
- 8723 Zhang, M., Eichhorn, S.W., Zingg, B., Yao, Z., Cotter, K., Zeng, H., Dong, H. and Zhuang, X.
 8724 (2021a). Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH.
 8725 *Nature* 598, 137-143.
- 8726 Zhang, M., Hu, S., Min, M., Ni, Y., Lu, Z., Sun, X., Wu, J., Liu, B., Ying, X. and Liu, Y. (2021b).
 8727 Dissecting transcriptional heterogeneity in primary gastric adenocarcinoma by single cell
 8728 RNA sequencing. *Gut* 70, 464-475.
- 8729 Zhang, M., Liu, S., Miao, Z., Han, F., Gottardo, R. and Sun, W. (2022d). IDEAS: individual level
 8730 differential expression analysis for single-cell RNA-seq data. *Genome Biol* 23, 33.
- 8731 Zhang, R., Meng-Papaxanthos, L., Vert, J.-P. and Noble, W.S. (2021c). Semi-supervised
 8732 single-cell cross-modality translation using Polarbear. *bioRxiv*, doi:
 8733 10.1101/2021.11.18.467517.
- 8734 Zhang, R., Zhou, T. and Ma, J. (2022e). Multiscale and integrative single-cell Hi-C analysis with
 8735 Higashi. *Nat Biotechnol* 40, 254-261.
- 8736 Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., *et al.*
 8737 (2019c). CellMarker: a manually curated resource of cell markers in human and mouse.

- 8738 Nucleic Acids Res 47, D721-d728.
- 8739 Zhang, X.F., Ou-Yang, L., Yang, S., Zhao, X.M., Hu, X. and Yan, H. (2019d). EnImpute: imputing
8740 dropout events in single-cell RNA-sequencing data via ensemble learning. Bioinformatics 35,
8741 4827-4829.
- 8742 Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.C. and Yates, J.R., 3rd. (2013). Protein analysis by
8743 shotgun/bottom-up proteomics. Chem Rev 113, 2343-2394.
- 8744 Zhang, Y., Li, Q., Jiang, N., Su, Z., Yuan, Q., Lv, L., Sang, X., Chen, R., Feng, Y. and Chen, Q.
8745 (2022f). Dihydroartemisinin beneficially regulates splenic immune cell heterogeneity through
8746 the SOD3-JNK-AP-1 axis. Sci China Life Sci 65, 1636-1654.
- 8747 Zhang, Y., Liu, T., Hu, X., Wang, M., Wang, J., Zou, B., Tan, P., Cui, T., Dou, Y., Ning, L., *et al.*
8748 (2021d). CellCall: integrating paired ligand-receptor and transcription factor activities for
8749 cell-cell communication. Nucleic Acids Res 49, 8520-8534.
- 8750 Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C.,
8751 Myers, R.M., Brown, M., Li, W., *et al.* (2008). Model-based analysis of ChIP-Seq (MACS).
8752 Genome Biol 9, R137.
- 8753 Zhang, Y. and Wang, F. (2021). SSBER: removing batch effect for single-cell RNA sequencing
8754 data. BMC Bioinformatics 22, 249.
- 8755 Zhang, Y., Xie, X., Wu, P. and Zhu, P. (2021e). SIEVE: identifying robust single cell variable
8756 genes for single-cell RNA sequencing data. Blood Sci 3, 35-39.
- 8757 Zhang, Y., Zhang, F., Wang, Z., Wu, S. and Tian, W. (2022g). scMAGIC: accurately annotating
8758 single cells using two rounds of reference-based classification. Nucleic Acids Res 50, e43.
- 8759 Zhao, C., Hu, S., Huo, X. and Zhang, Y. (2017). Dr.seq2: A quality control and analysis pipeline
8760 for parallel single cell transcriptome and epigenome data. PLoS One 12, e0180583.
- 8761 Zhao, E., Stone, M.R., Ren, X., Guenthoer, J., Smythe, K.S., Pulliam, T., Williams, S.R., Uytingco,
8762 C.R., Taylor, S.E.B., Nghiem, P., *et al.* (2021a). Spatial transcriptomics at subspot resolution
8763 with BayesSpace. Nat Biotechnol 39, 1375-1384.
- 8764 Zhao, J., Wang, G., Ming, J., Lin, Z., Wang, Y., Wu, A.R. and Yang, C. (2022a). Adversarial
8765 domain translation networks for fast and accurate integration of large-scale atlas-level
8766 single-cell datasets. bioRxiv, doi: 10.1101/2021.11.16.468892.
- 8767 Zhao, M., Jiang, J., Zhao, M., Chang, C., Wu, H. and Lu, Q. (2021b). The Application of
8768 Single-Cell RNA Sequencing in Studies of Autoimmune Diseases: a Comprehensive Review.
8769 Clin Rev Allergy Immunol 60, 68-86.
- 8770 Zhao, T., Chiang, Z.D., Morriss, J.W., LaFave, L.M., Murray, E.M., Del Priore, I., Meli, K.,
8771 Lareau, C.A., Nadaf, N.M., Li, J., *et al.* (2022b). Spatial genomics enables multi-modal study
8772 of clonal heterogeneity in tissues. Nature 601, 85-91.
- 8773 Zhao, X., Guo, J., Nie, F., Chen, L., Li, Z. and Zhang, H. (2020). Joint Principal Component and
8774 Discriminant Analysis for Dimensionality Reduction. IEEE Trans Neural Netw Learn Syst 31,
8775 433-444.
- 8776 Zhao, Y., Wang, T., Liu, Z., Ke, Y., Li, R., Chen, H., You, Y., Wu, G., Cao, S., Du, Z., *et al.* (2023).
8777 Single-cell transcriptomics of immune cells in lymph nodes reveals their composition and
8778 alterations in functional dynamics during the early stages of bubonic plague. Sci China Life
8779 Sci 66, 110-126.
- 8780 Zhao, Z., Zhu, H., Li, Q., Liao, W., Chen, K., Yang, M., Long, D., He, Z., Zhao, M., Wu, H., *et al.*
8781 (2022c). Skin CD4(+) Trm cells distinguish acute cutaneous lupus erythematosus from

- 8782 localized discoid lupus erythematosus/subacute cutaneous lupus erythematosus and other
 8783 skin diseases. *J Autoimmun* 128, 102811.
- 8784 Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler,
 8785 T.D., McDermott, G.P., Zhu, J., *et al.* (2017). Massively parallel digital transcriptional
 8786 profiling of single cells. *Nat Commun* 8, 14049.
- 8787 Zheng, M., Hu, Z., Mei, X., Ouyang, L., Song, Y., Zhou, W., Kong, Y., Wu, R., Rao, S., Long, H.,
 8788 *et al.* (2022a). Single-cell sequencing shows cellular heterogeneity of cutaneous lesions in
 8789 lupus erythematosus. *Nat Commun* 13, 7489.
- 8790 Zheng, R., Dong, X., Wan, C., Shi, X., Zhang, X. and Meyer, C.A. (2020). Cistrome Data Browser
 8791 and Toolkit: analyzing human and mouse genomic data using compendia of ChIP-seq and
 8792 chromatin accessibility data. *Quantitative Biology* 8, 267-276.
- 8793 Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C.H., Brown, M., Zhang, X., Meyer,
 8794 C.A., *et al.* (2019). Cistrome Data Browser: expanded datasets and new tools for gene
 8795 regulatory analysis. *Nucleic Acids Res* 47, D729-d735.
- 8796 Zheng, R., Zhang, Y., Tsuji, T., Gao, X., Wagner, A., Yosef, N., Chen, H., Zhang, L., Tseng, Y.-H.
 8797 and Chen, K. (2022b). MEBOCOST: Metabolite-mediated Cell Communication Modeling by
 8798 Single Cell Transcriptome. *bioRxiv*, doi: 10.1101/2022.05.30.494067.
- 8799 Zheng, X.T. and Li, C.M. (2012). Single cell analysis at the nanoscale. *Chem Soc Rev* 41,
 8800 2061-2071.
- 8801 Zhong, L., Yang, X., Zhou, Y., Xiao, J., Li, H., Tao, J., Xi, Q., Chu, C., Li, C., Yang, X., *et al.*
 8802 (2022). Exploring the R-ISS stage-specific regular networks in the progression of multiple
 8803 myeloma at single-cell resolution. *Sci China Life Sci* 65, 1811-1823.
- 8804 Zhou, J., Ma, J., Chen, Y., Cheng, C., Bao, B., Peng, J., Sejnowski, T.J., Dixon, J.R. and Ecker, J.R.
 8805 (2019a). Robust single-cell Hi-C clustering by convolution- and random-walk-based
 8806 imputation. *Proc Natl Acad Sci U S A* 116, 14011-14018.
- 8807 Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C. and
 8808 Chanda, S.K. (2019b). Metascape provides a biologist-oriented resource for the analysis of
 8809 systems-level datasets. *Nat Commun* 10, 1523.
- 8810 Zhu, A., Srivastava, A., Ibrahim, J.G., Patro, R. and Love, M.I. (2019a). Nonparametric expression
 8811 analysis using inferential replicate counts. *Nucleic Acids Res* 47, e105.
- 8812 Zhu, C., Yu, M., Huang, H., Juric, I., Abnousi, A., Hu, R., Lucero, J., Behrens, M.M., Hu, M. and
 8813 Ren, B. (2019b). An ultra high-throughput method for single-cell joint analysis of open
 8814 chromatin and transcriptome. *Nat Struct Mol Biol* 26, 1063-1070.
- 8815 Zhu, C., Zhang, Y., Li, Y.E., Lucero, J., Behrens, M.M. and Ren, B. (2021a). Joint profiling of
 8816 histone modifications and transcriptome in single cells from mouse brain. *Nat Methods* 18,
 8817 283-292.
- 8818 Zhu, H. and Wang, Z. (2019). SCL: a lattice-based approach to infer 3D chromosome structures
 8819 from single-cell Hi-C data. *Bioinformatics* 35, 3981-3988.
- 8820 Zhu, H., Zou, G., Wang, N., Zhuang, M., Xiong, W. and Huang, G. (2017). Single-neuron
 8821 identification of chemical constituents, physiological changes, and metabolism using mass
 8822 spectrometry. *Proc Natl Acad Sci U S A* 114, 2586-2591.
- 8823 Zhu, J. and Sabatti, C. (2020). Integrative Spatial Single-cell Analysis with Graph-based Feature
 8824 Learning. *bioRxiv*, doi: 10.1101/2020.08.12.248971.
- 8825 Zhu, J., Sun, S. and Zhou, X. (2021b). SPARK-X: non-parametric modeling enables scalable and

- robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biol* 22, 184.
- Zhu, P., Guo, H., Ren, Y., Hou, Y., Dong, J., Li, R., Lian, Y., Fan, X., Hu, B., Gao, Y., *et al.* (2018a). Single-cell DNA methylome sequencing of human preimplantation embryos. *Nat Genet* 50, 12-19.
- Zhu, Q., Shah, S., Dries, R., Cai, L. and Yuan, G.C. (2018b). Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence *in situ* hybridization data. *Nat Biotechnol*, doi: 10.1038/nbt.4260.
- Zhu, Y., Clair, G., Chrisler, W.B., Shen, Y., Zhao, R., Shukla, A.K., Moore, R.J., Misra, R.S., Pryhuber, G.S., Smith, R.D., *et al.* (2018c). Proteomic Analysis of Single Mammalian Cells Enabled by Microfluidic Nanodroplet Sample Preparation and Ultrasensitive NanoLC-MS. *Angew Chemie Int Ed Engl* 57, 12370-12374.
- Zhu, Y., Piehowski, P.D., Zhao, R., Chen, J., Shen, Y., Moore, R.J., Shukla, A.K., Petyuk, V.A., Campbell-Thompson, M., Mathews, C.E., *et al.* (2018d). Nanodroplet processing platform for deep and quantitative proteome profiling of 10-100 mammalian cells. *Nat Commun* 9, 882.
- Zhu, Y., Podolak, J., Zhao, R., Shukla, A.K., Moore, R.J., Thomas, G.V. and Kelly, R.T. (2018e). Proteome Profiling of 1 to 5 Spiked Circulating Tumor Cells Isolated from Whole Blood Using Immunodensity Enrichment, Laser Capture Microdissection, Nanodroplet Sample Processing, and Ultrasensitive nanoLC-MS. *Anal Chem* 90, 11756-11759.
- Zhu, Y., Zhang, Y.-X., Cai, L.-F. and Fang, Q. (2013). Sequential Operation Droplet Array: An Automated Microfluidic Platform for Picoliter-Scale Liquid Handling, Analysis, and Screening. *Anal Chem* 85, 6723-6731.
- Zhu, Y., Zhao, R., Piehowski, P.D., Moore, R.J., Lim, S., Orphan, V.J., Paša-Tolić, L., Qian, W.J., Smith, R.D. and Kelly, R.T. (2018f). Subnanogram proteomics: impact of LC column selection, MS instrumentation and data analysis strategy on proteome coverage for trace samples. *Int J Mass spectrom* 427, 4-10.
- Zhuang, X. (2021). Spatially resolved single-cell genomics and transcriptomics by imaging. *Nat Methods* 18, 18-22.
- Zong, C., Lu, S., Chapman, A.R. and Xie, X.S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 338, 1622-1626.
- Zong, Y., Yu, T., Wang, X., Wang, Y., Hu, Z. and Li, Y. (2022). conST: an interpretable multi-modal contrastive learning framework for spatial transcriptomics. *bioRxiv*, doi: 10.1101/2022.01.14.476408.
- Zou, H. and Tibshirani, H.R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Stats* 15, 1-30.
- Zou, J., Deng, F., Wang, M., Zhang, Z., Liu, Z., Zhang, X., Hua, R., Chen, K., Zou, X. and Hao, J. (2022). scCODE: an R package for data-specific differentially expressed gene detection on single-cell RNA-sequencing data. *Brief Bioinform* 23, bbac180.
- Zou, Z., Hua, K. and Zhang, X. (2021). HGC: fast hierarchical clustering for large-scale single-cell data. *Bioinformatics* 37, 3964-3965.
- Zuo, C., Zhang, Y., Cao, C., Feng, J., Jiao, M. and Chen, L. (2022). Elucidating tumor heterogeneity from spatially resolved transcriptomics data by multi-view graph collaborative

- 8870 learning. *Nat Commun* 13, 5962.
- 8871 Žurauskienė, J. and Yau, C. (2016). *pcaReduce: hierarchical clustering of single cell*
- 8872 *transcriptional profiles. BMC Bioinformatics* 17, 140.
- 8873

Accepted

Table1 Comparison of mainstream WGA methods

WGA method	Amplification principle	Enzyme	Product length	Coverage	Uniformity	Accuracy	Application	Reference
PEP-PCR	PCR-based	DNA polymerase	<2 kb	~50%	+++	+	CGH, LOH, STR, etc.	(Zhang et al., 1992)
DOP-PCR	PCR-based	DNA polymerase	<2 kb	~45%	+++	+	FISH, SNP, SSCP, etc.	(Telenius et al., 1992)
MDA	Isothermal amplification	phi29 DNA polymerase	<100 kb	~87%	++	+++	NGS, SNV, SNP, STR, single-cell sequencing, etc.	(Dean et al., 2001)
eMDA	Isothermal amplification	phi29 DNA polymerase	<100 kb	72%	+++	+++	CNV, SNV, single-cell sequencing, etc.	(Fu et al., 2015)
SISSOR	Isothermal amplification	phi29 DNA polymerase	<100 kb	~70%	+++	++++	single-cell sequencing, haploid analysis, etc.	(Chu et al., 2017)
PTA	Isothermal amplification	phi29 DNA polymerase	~150 bp	>95%	+++++	+++++	CNV, SNV, single-cell sequencing, etc.	(Gonzalez-Pena et al., 2021)
MALBAC	Combining isothermal amplification with PCR	Bst enzyme; Taq DNA polymerase	<2 kb	~93%	+++	++	single-cell sequencing, NGS, STR, CGH, SNV, CNV, etc.	(Zong et al., 2012)
LIANTI	Linear amplification	Reverse transcriptase; T7 RNA polymerase	~400 bp	~97%	++++	+++	CNV, SNV, CGH, single-cell sequencing, chromosome analysis, etc.	(Chen et al., 2017a)
META-CS	End-tagging-based PCR amplification	Q5 DNA polymerase	<2 kb	64%	+++	++++	SNV, single-cell sequencing, etc.	(Xing et al., 2021)

CGH: comparative genomic hybridization; LOH: loss of heterozygosity; STR: short sequence repeat; FISH: fluorescence in situ hybridization; SNP: single nucleotide polymorphism; SSCP: single-strand conformational polymorphism; NGS: second generation sequencing; CNV: copy number variation; SNV: single nucleotide variation

Table 2 Recent SCP tools

SCP Tools	Customized Equipment	Label	Cell Type	Isolation Method	MS Approach	Pretreatment Throughput	MS Throughput	Average Identified Protein Number per Cell	Depth of Proteome Coverage	Reference
iPAD-1	Need	label free	HeLa	capillary-based isolation	LC-MS/MS	1 cell per run	24 cells per day	271	406 (n = 10)	(Shao et al., 2018)
OAD	Need	label free	HeLa	SODA	LC-MS/MS	1 cell per run	4 cells per day	51	/	(Li et al., 2018b)
	Need	label free	HeLa	FACS	LC-MS/MS	27 cells per run	/	669	/	(Zhu et al., 2018c)
nanoPOTS	Need	label free	MCF10A				24 cells per day	764	1093 (n = 10)	
	Need	TMT10-plex labelled	MOLM-14, K562, CMK	FACS	LC-MS/MS	/	77 cells per day	1281	2558 (n = 152)	(Williams et al., 2020)
autoPOTS	No need	label free	HeLa	FACS	LC-MS/MS	/	10 cells per day	301	/	(Liang et al., 2021c)
nested nanoPOTS	Need	TMT16-plex labelled	C10, RAW, SVEC	cellenONE	LC-MS/MS	243 cells per run	108 cells per day	1716	2457 (n = 108)	(Woo et al., 2021)
proteoCHIP	Need	TMT16-plex labelled	HeLa, HEK-293	cellenONE	LC-FAIMS-MS/MS	592 cells per run	384 cells per day	1940 (20x carrier) 1598 (no-carrier)	3674 (n = 276)	(Ctortecka et al., 2022a)
iProChip	Need	label free	MEC-1	chip device	LC-MS/MS	9 cell per run	9 cells per day	455	/	
SciProChip	Need	label free	PC-9	/	/	20 cell per run	16 cells per day	1500	1995 (n = 10)	(Gebreyesus et al., 2022)
PiSPA	Need	label free	A549	SODA	LC-TIMS-TOF	1 cell per run	/	3008	5093 (n = 37)	(Wang et al., 2022d)
SCoPE	No need	TMT10-plex labelled	Jurkat, U-937	manual picking	LC-MS/MS	8 cells per run	48 cells per day	/	767 (n = 24)	(Budnik et al., 2018)
SCoPE2	No need	TMT16-plex labelled	monocyte and macrophage cells	FACS	LC-MS/MS	/	200 cells per day	1000	3042 (n = 1490)	(Specht et al., 2021)
nPOP-SCoPE 2	Need	TMT18-plex labelled	U-937, WM989	cellenONE	LC-MS/MS	2016 cells per run	212 cells per day	997	2844 (n = 1543)	(Leduc et al., 2022)
A multiplexed scMS workflow	No need	TMT16-plex labelled	OCI-AML8227	FACS	LC-FAIMS-MS/MS	336 cells per run	112 cells per day	987	2723 (n = 2050)	(Schoof et al., 2021)
UE-SCP	No need	TMT6-plex labelled	HeLa, HEK-293T	cellenONE	LC-TIMS-TOF	308 cells per run	96 cells per day	2249	4230 (n = 128)	(Gu et al., 2022b)
Mad-CASP	No need	label free	HeLa	FACS	LC-MS/MS	/	16 cells per day	1240	/	(Li et al., 2022g)
WinO	No need	TMT10-plex labelled	RPMI8226	SH800S Cell sorter	LC-MS/MS	/	144 cells per day	845	/	(Masuda et al., 2022)
T-SCP	No need	label free	HeLa	FACS	LC-TIMS-TOF-SCP	308 cells per run	41 cells per day	2083	2501 (n = 231)	(Brunner et al., 2022)

Table 3 Recent single-cell multimodal technologies

Multimodal techniques	Modalities	References
Genome and transcriptome sequencing (G&T-seq)	gDNA+mRNA	(Macaulay et al., 2015)
gDNA-mRNA sequencing (DR-seq)	gDNA+mRNA	(Dey et al., 2015)
Simultaneous isolation of genomic DNA and total RNA (SIDR)	gDNA+mRNA	(Han et al., 2018a)
TARGET-seq	gDNA+mRNA	(Rodriguez-Meira et al., 2019)
single-cell methylome and transcriptome sequencing (scM&T-seq)	mRNA+Methylation	(Angermueller et al., 2016)
Simultaneous single-cell methylome and transcriptome sequencing (scMT-seq)	mRNA+Methylation	(Hu et al., 2016b)
scTrio-Seq	mRNA+Methylation	(Hou et al., 2016)
sci-CAR	mRNA+ATAC	(Cao et al., 2018)
SNARE-seq	mRNA+ATAC	(Chen et al., 2019d)
Paired-seq	mRNA+ATAC	(Zhu et al., 2019b)
SHARE-seq	mRNA+ATAC	(Ma et al., 2020)
10x Multiome	mRNA+ATAC	https://www.10xgenomics.com/cn /blog/introducing-chromium-single-cell-multiome-atac-gene-expression
PEA/STA	mRNA+proteome	(Genshaft et al., 2016)
PLAYR	mRNA+proteome	(Frei et al., 2016)
CITE-seq	mRNA+proteome	(Stoeckius et al., 2017)
REAP-seq	mRNA+proteome	(Peterson et al., 2017)
RAID	mRNA+proteome	(Gerlach et al., 2019)
scNMT-seq	mRNA+methylation+ATAC	(Clark et al., 2018)
scNOMeRe-seq	mRNA+methylation+ATAC	(Wang et al., 2021b)
ECCITE-seq	mRNA+sgRNA+target protein	(Mimitou et al., 2019)
Paired-Tag	mRNA+ATAC+5 histone modifications	(Zhu et al., 2021a)
scCUT&Tag-pro	5 histone modifications+proteome	(Zhang et al., 2022a)

Table 4 The curated list of spatial transcriptomics techniques and their corresponding features

Techniques	Features	Type	Spatial resolution	Gene coverage	References
tomo-seq	Require identical biological samples; Enable the 3D reconstruction of tissues	Microdissection-based	18 μ m	Transcriptome-wide	(Junker et al., 2014)
STRP-seq	Require consecutive thin slices	Microdissection-based	10 cells	Transcriptome-wide	(Schede et al., 2021)
Geo-seq	Enable the 3D reconstruction of tissues	Microdissection-based	10 cells	Transcriptome-wide	(Chen et al., 2017b)
PIC-seq	Focus on physical cell interaction rather than spatial positions	Microdissection-based	Cellular	Transcriptome-wide	(Giladi et al., 2020)
ProximID	Focus on physical cell interaction rather than spatial positions	Microdissection-based	Cellular	Transcriptome-wide	(Boisset et al., 2018)
TIVA	Require the loading of capture tag into cells; Rely on photoactivation	Microdissection-based	Cellular	Transcriptome-wide	(Lovatt et al., 2014)
NICHE-seq	Work on genetically engineered mice; Select region of interest by photoactivation	Microdissection-based	Cellula	Transcriptome-wide	(Medaglia et al., 2017)
GeomX DSP	Rely on photocleavage; Applicable in protein detection	Microdissection-based	20~40 cells	Targeted (~1,500 genes)	(Merritt et al., 2020)
ST	With H&E image	Spatial barcoding	100 μ m	Transcriptome-wide	(Stahl et al., 2016)
10x Visium	With H&E or immunohistochemistry image	Spatial barcoding	55 μ m	Transcriptome-wide	(Stahl et al., 2016)
Slide-seq(V2)	Without histology on the same tissue section	Spatial barcoding	10 μ m	Transcriptome-wide	(Stickels et al., 2021)
HDST	Low sensitivity	Spatial barcoding	2 μ m	Transcriptome-wide	(Vickovic et al., 2019)
Stereo-seq	Enable subcellular analysis; Allow large field of view	Spatial barcoding	0.5~0.7 μ m	Transcriptome-wide	(Chen et al., 2022)
Seq-scope	Enable subcellular analysis	Spatial barcodin	0.5~0.8 μ m	Transcriptome-wide	(Cho et al., 2021)
PIXEL-seq	Enable subcellular analysis	Spatial barcodin	1 μ m	Transcriptome-wide	(Fu et al., 2021b)
DBiT-seq	Enable simultaneous measurement of RNA and proteins	Spatial barcodin	10 μ m	Transcriptome-wide	(Liu et al., 2020b)
seqFISH+	Use a barcode palette of 60 pseudo colours	Fluorescence imaging (ISH-based)	Subcellular	Targeted (~10,000 genes)	(Eng et al., 2019)
MERFISH	Based on multi-bit binary encoding strategy; Combine expansion microscopy with <i>in situ</i> hybridization	Fluorescence imaging (ISH-based)	Subcellular	Targeted (~10,000 genes)	(Xia et al., 2019)
STARmap	Combine hydrogel-tissue chemistry with <i>in situ</i> sequencing	Fluorescence imaging (ISS-based)	Subcellular	Targeted (~1,000 genes)	(Wang et al., 2018)
FISSEQ	Use partition sequencing	Fluorescence imaging (ISS-based)	Subcellular	Transcriptome-wide	(Lee et al., 2015)
ExSeq	Combine expansion microscopy with <i>in situ</i> sequencing	Fluorescence imaging (ISS-based)	Subcellular	Transcriptome-wide	(Shah et al., 2016; Alon et al., 2021)
Slide-DNA-seq	Use barcoded bead arrays to capture spatially resolved DNA	Spatial barcoding	10 μ m	Genome-wide	(Zhao et al., 2022b)
spatial-CUT&Tag	Combine <i>in situ</i> CUT&Tag chemistry with microfluidic deterministic barcoding	Spatial barcoding	20 μ m	Genome-wide	(Deng et al., 2022a)
spatial-ATAC-seq	Combine <i>in situ</i> Tn5 transposition chemistry and microfluidic deterministic barcoding	Spatial barcoding	20 μ m	Genome-wide	(Deng et al., 2022b)
DNA-MERFISH	Enable simultaneous imaging of genomic loci and nascent transcripts	Fluorescence imaging	Subcellular	Genome-wide	(Su et al., 2020)
CosMX SMI	Enable quantification of RNA and proteins	Fluorescence imaging	Subcellular	64 proteins, 1,000 genes	(He et al., 2021a)
SpaceM	Combine light microscopy and MALDI-imaging MS	Fluorescence imaging	cellular	>100 metabolites	(Rappez et al., 2021)
Perturb-map	Combine a protein barcode system and multiplex imaging	Fluorescence imaging	cellular	35 genes (120 Pro-Codes)	(Dhainaut et al., 2022)

Table 5 The different scCRISPR-seq platforms

Platforms	Omics	In vivo/vitro	Subject	sgRNA sequenced directly without barcode	No. of delivery of sgRNAs	Coupled CRISPR system	References
Perturb-Seq	Transcriptome	In vitro	K562	no	Single	CRISPR knockout	(Dixit et al., 2016)
Perturb-seq	Transcriptome	In vitro	K562	no	Single	CRISPR interference	(Adamson et al., 2016)
CRISP-seq	Transcriptome	In vitro	Mouse BMDCs	no	Single	CRISPR knockout	(Jaitin et al., 2016)
CROP-seq	Transcriptome	In vitro	Jurkat	yes	Paired	CRISPR knockout	(Datlinger et al., 2017)
Mosaic-seq	Multimodal (Transcriptome, Enhancer)	In vitro	K562	no	Single	CRISPR interference	(Xie et al., 2017)
Improved CROP-seq	Transcriptome	In vitro	MCF10A	yes	Paired	CRISPR knockout and interference	(Hill et al., 2018)
Perturb-ATAC	Epigenome (Chromatin accessibility)	In vitro	Primary human keratinocytes, B lymphoblasts	no	Single	CRISPR knockout and interference	(Rubin et al., 2019)
ECCITE-seq	Multimodal (Transcriptome, proteome, clonotypes)	In vitro	Sez4, MyLa, PBMC, NIH-3T3	yes	Single	CRISPR knockout	(Mimitou et al., 2019)
Direct-capture Perturb-seq	Transcriptome	In vitro	iPSCs, K562	yes	Multiple	CRISPR knockout, interference and activation	(Replogle et al., 2020)
Direct-seq	Transcriptome	In vitro	Jurkat, K562	yes	Multiple	CRISPR knockout and activation	(Song et al., 2020)
In vivo Perturb-seq	Transcriptome	In vivo	Progenitor cells of the mouse forebrain	no	Single	CRISPR interference	(Jin et al., 2020b)
PoKI-seq	Transcriptome	In vivo	Human primary T cells, NSG mice bearing human melanoma cells	no	Single	CRISPR knock-in	(Roth et al., 2020)
Spear-ATAC	Epigenome (Chromatin accessibility)	In vitro	K562, GM12878, MCF7	yes	Single	CRISPR knockout and interference	(Pierce et al., 2021)
CRISPR-sciATAC	Epigenome (Chromatin accessibility)	In vitro	NIH-3T3, K562	yes	Single	CRISPR knockout	(Liscovitch-Brauer et al., 2021)
genome-scale Perturb-seq	Transcriptome	In vitro	K562, RPE1	yes	Single	CRISPR interference	(Replogle et al., 2022)
Perturb-map	Multimodal (Spatial transcriptome, imaging)	In vitro, in vivo	293T, KP, 4T1	no	Single	CRISPR knockout	(Dhainaut et al., 2022)
AAV-Perturb-seq	Transcriptome	In vivo	Adult mouse brain prefrontal cortex	yes	Single	CRISPR interference	(Santinha et al., 2023)
CHOOSE screen	Transcriptome	In vitro (organoid)	Brain organoid	no	Paired	CRISPR knockout	(Li et al., 2023)

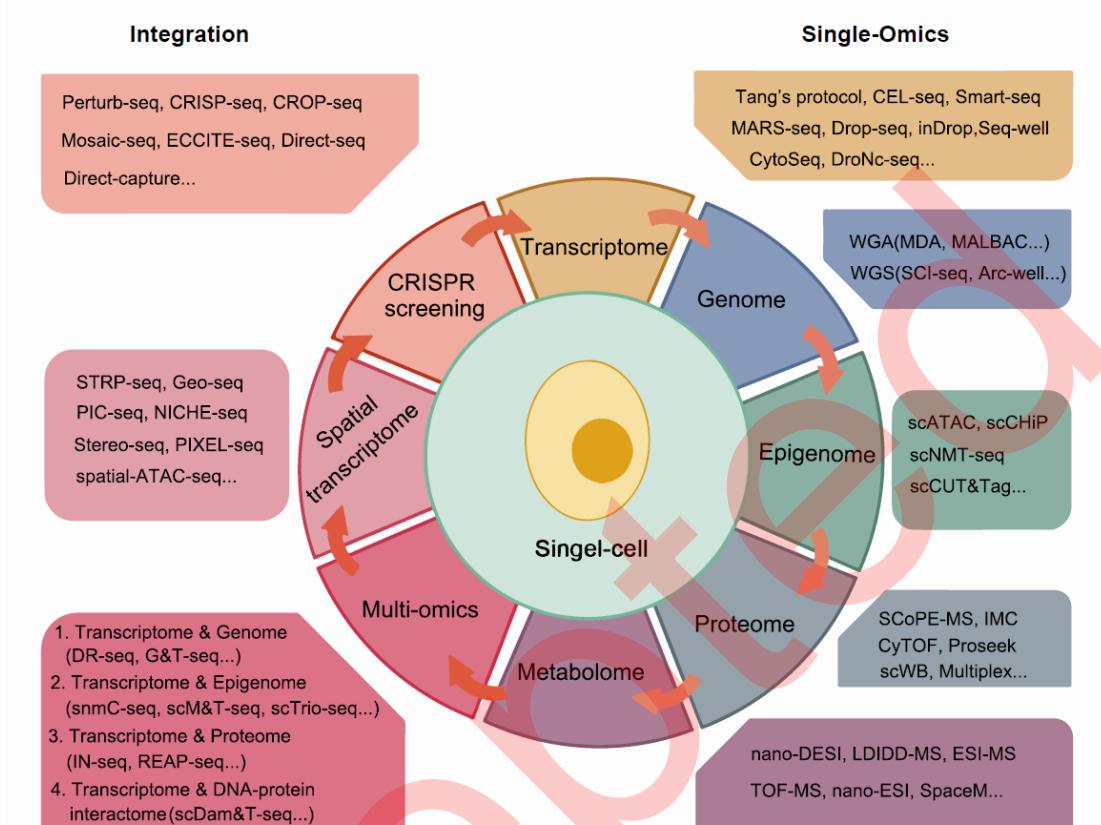


Figure 1 Schematic diagram of single-cell sequencing technologies. Since the inception of the first scRNA-seq in 2009, single-cell sequencing technology has been rapidly expanded to other omics levels and diverse integration approaches. Single omics-level sequencing technologies now include transcriptome, genome, epigenome, proteome and metabolome. Integrated sequencing technologies involve multiple omics data integration, such as transcriptome & genome, transcriptome & epigenome, transcriptome & proteome, transcriptome & DNA–protein interactome. Additionally, these integrated approaches incorporate sequencing data with other layers of information, including spatial data and the CRISPR screening technique. Each type of expansion is represented by the technology listed in the corresponding color box.

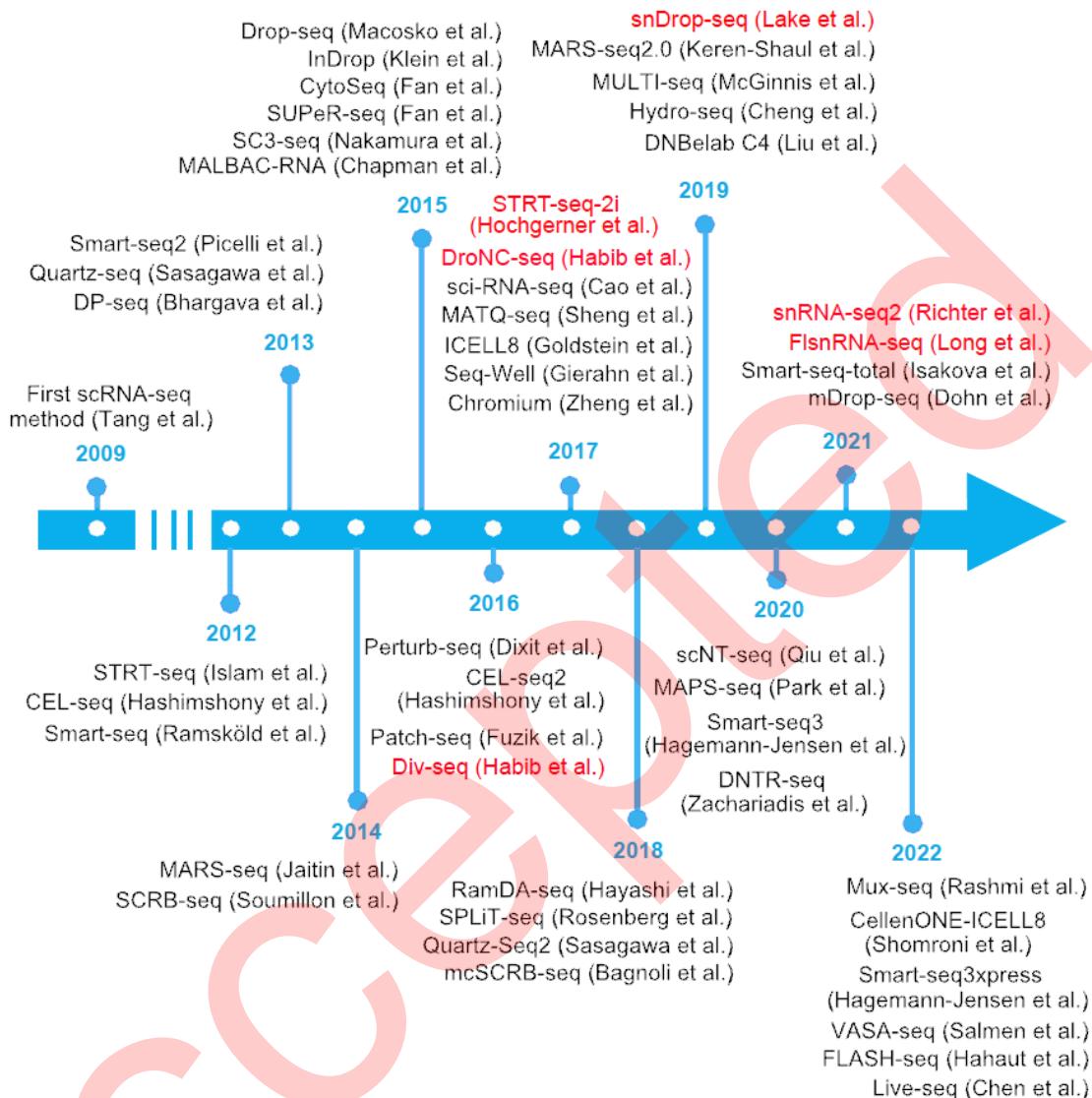


Figure 2 Significant works in the field of scRNA-seq over the past 10 years. Black represents scRNA-seq technologies; Red represents snRNA-seq technologies.

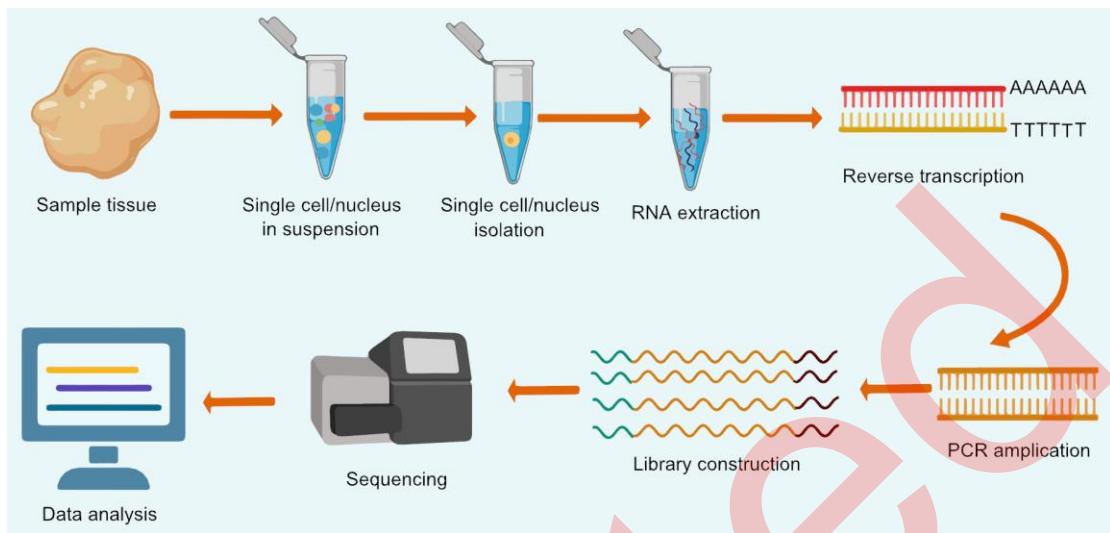


Figure 3 Major steps of scRNA-seq workflow.

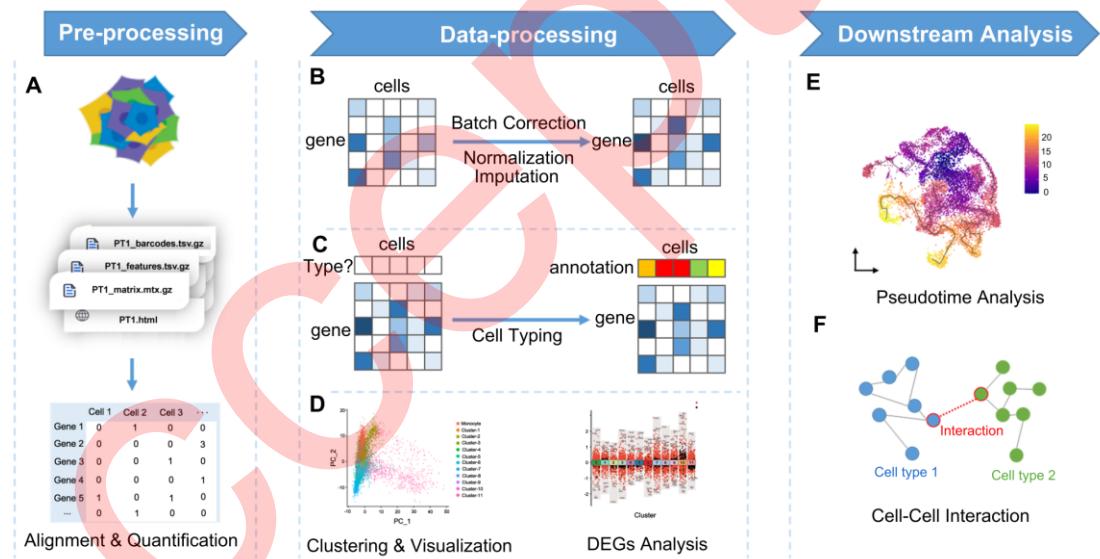


Figure 4 Overview of the single cell analysis workflow. A, The cell-gene matrix count is formed based on sequencing data through single cell read alignment and quantification methods in pre-processing stage. B, High quality cell matrix used for analysis is obtained by processing the original gene expression matrix, make batch correction to remove batch effect, normalize to reduce biological differences and fill in genes that were missed in sequencing. C, Make cell types assignment based on prior references or not. D, Cells with similar transcriptome characteristics are grouped into one cell group called cluster and visualization of cells can be realized by the method of dimension reduction. Differential Expression Analysis (DEGs) check the significance of the classification between groups. E, Pseudotime analysis can restore the dynamic process of cellular transcript change. F, Transcriptome regulatory relationships between cells can be inferred through cell-cell interaction analysis.

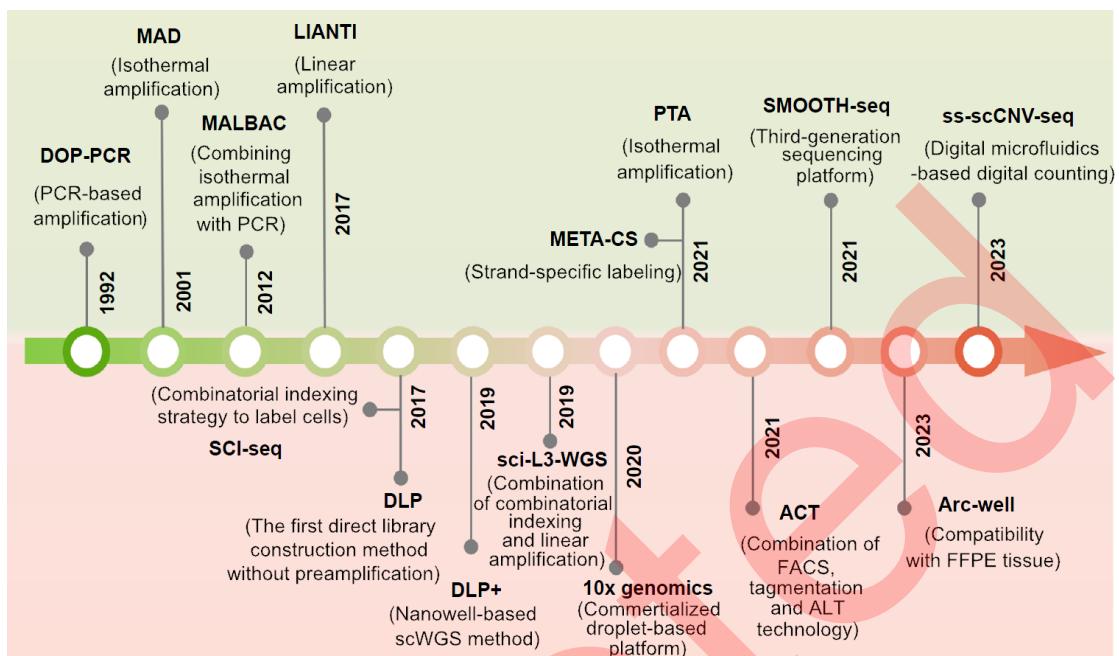


Figure 5. Timeline of the development of single-cell whole genome amplification and sequencing technology. The upper half of the graph shows the main events marking the development of scWGA and low-throughput scWGS technologies, and the bottom half the development in throughput.

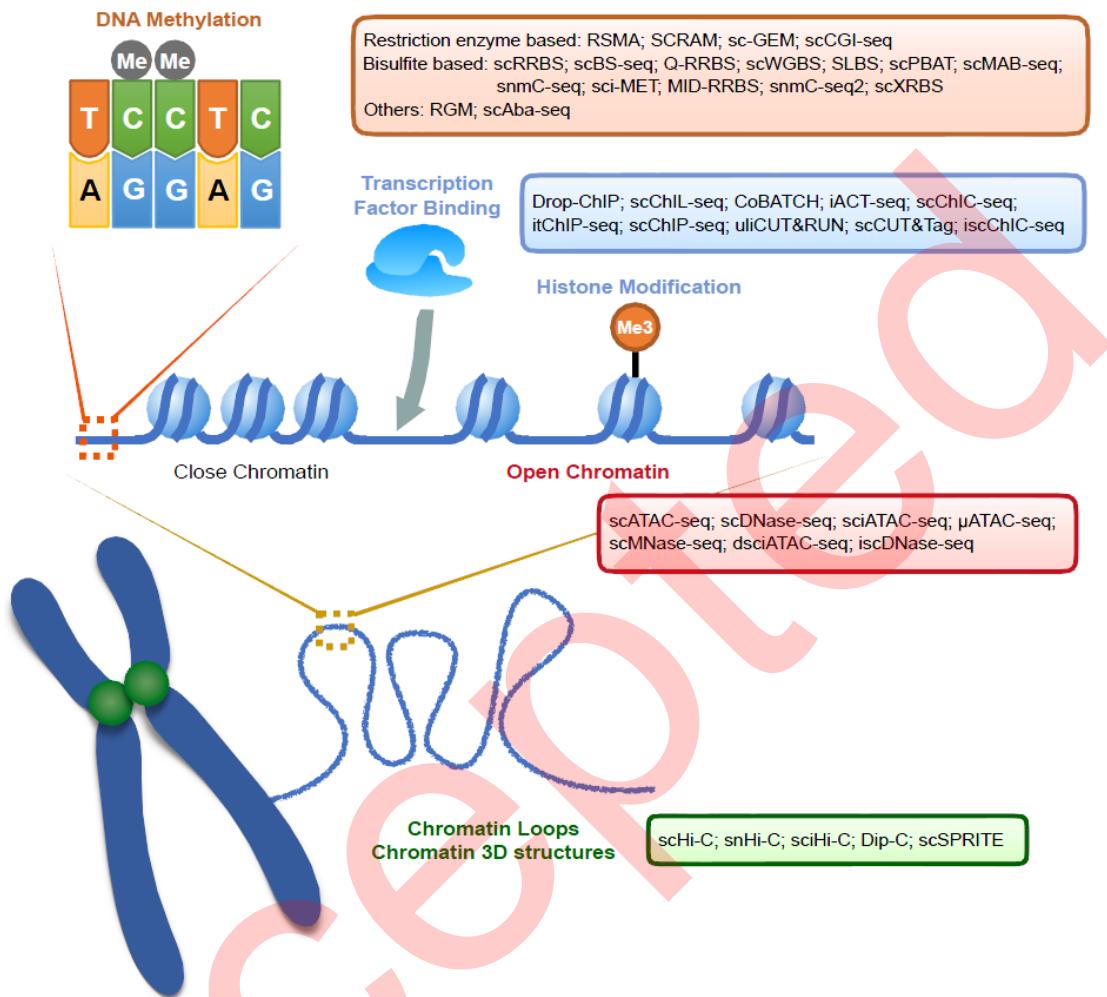


Figure 6 Overview of single-cell epigenome sequencing techniques. Single-cell epigenome sequencing is mainly performed at four levels, namely DNA modification, transcription factor binding or histone modification, chromatin accessibility and chromatin three-dimensional structure. DNA modification mainly refers to DNA methylation, and the sequencing methods for DNA modification are mainly based on restriction enzyme digestion and bisulfite conversion, which are shown in orange in the figure. The sequencing methods for transcription factor binding and histone modification are shown in blue. The sequencing methods for chromatin accessibility are shown in red. The sequencing methods for chromatin three-dimensional structure are shown in green.

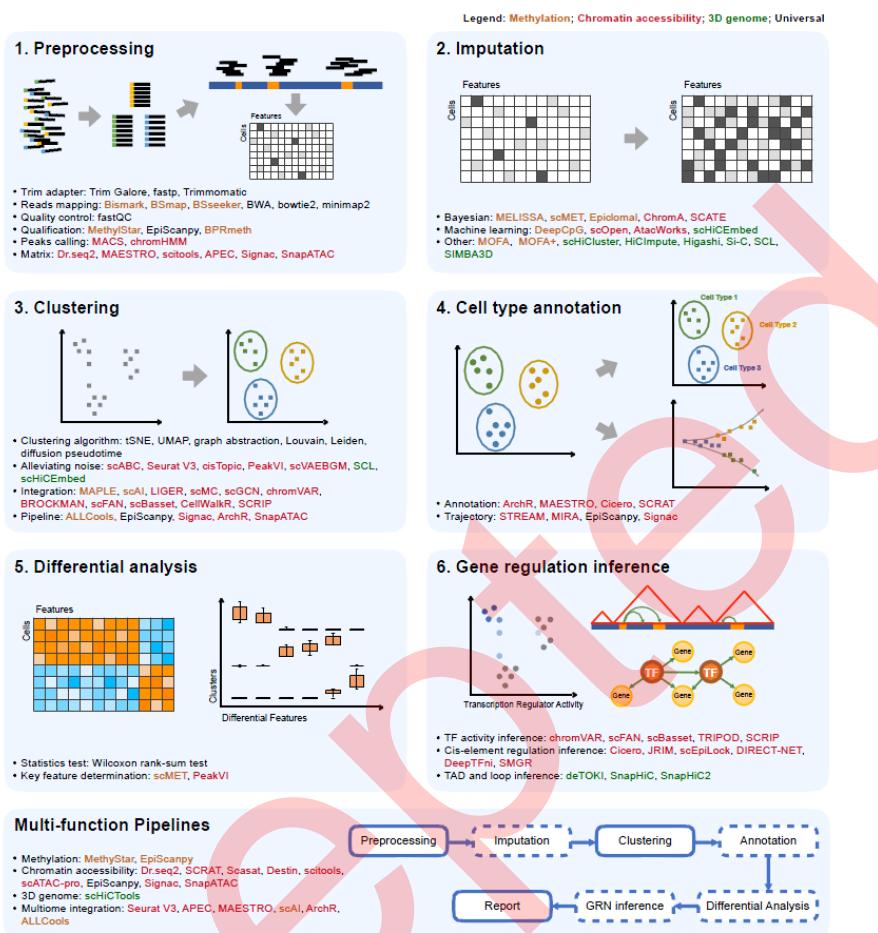


Figure 7 Overview of the major steps for single-cell epigenomic analysis and related computation methods. The analysis of single-cell epigenetic data can be roughly divided into 6 parts as shown in the figure. First, the data needs to be preprocessed. The preprocessing part includes removing sequencing adapters, sequence alignment, quality control, data quantification, peak calling and building feature matrix. Due to the extreme sparsity of single-cell epigenetic data, there are currently methods based on Bayesian inference and machine learning to impute missing data. After that, the data is used for dimensionality reduction and clustering. After clustering, different cell types are annotated and pseudotime analysis can be performed. Then, based on different cell types, different genomic features of different cells can be identified. For studying gene transcription regulation, transcription factor binding activity, cis-regulatory elements regulation and TAD regions and loops identification can be inferred. There are also pipelines that integrate multiple functions for batch processing data and output multiple results. Among them, preprocessing, clustering, reporting and other functions are common in all pipelines, while imputation, cell type annotation, differential analysis and gene regulatory network inference are optional functions in different pipelines. In the figure, methods marked in brown are specifically for processing DNA methylation data, methods marked in red are specifically for processing chromatin accessibility data, methods marked in green are specifically for processing three-dimensional genome data and methods marked in black are applicable to two or more types of data.

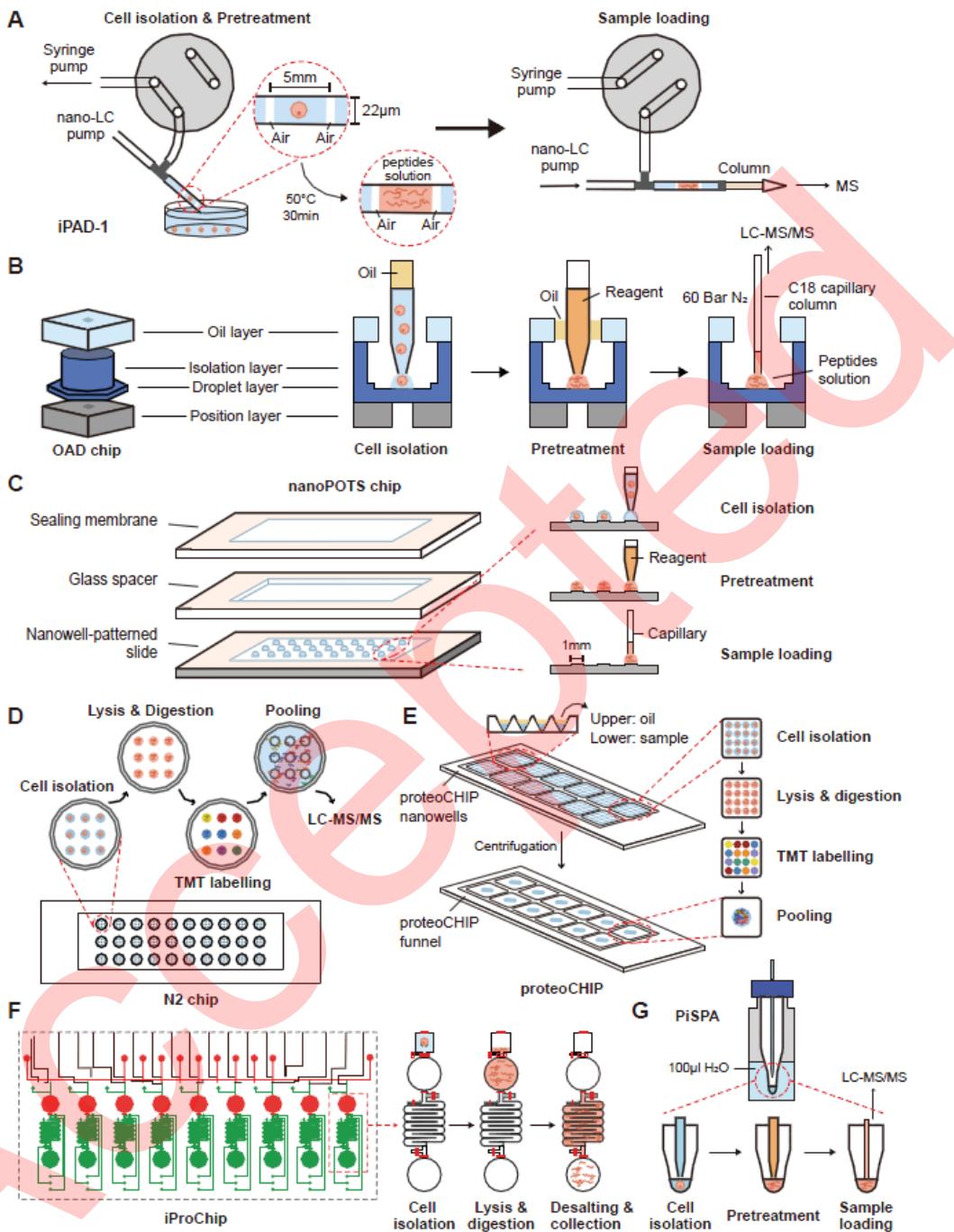


Figure 8 Workflows of recent SCP tools. A, iPAD-1 (Shao et al., 2018). B, OAD (Li et al., 2018c). C, nanoPOTS (Zhu et al., 2018d). D, nest nanoPOTS (Woo et al., 2021). E, proteoCHIP (Ctortecka et al., 2022a). F, iProChip (Gebreyesus et al., 2022). G, PiSPA (Wang et al., 2022d).

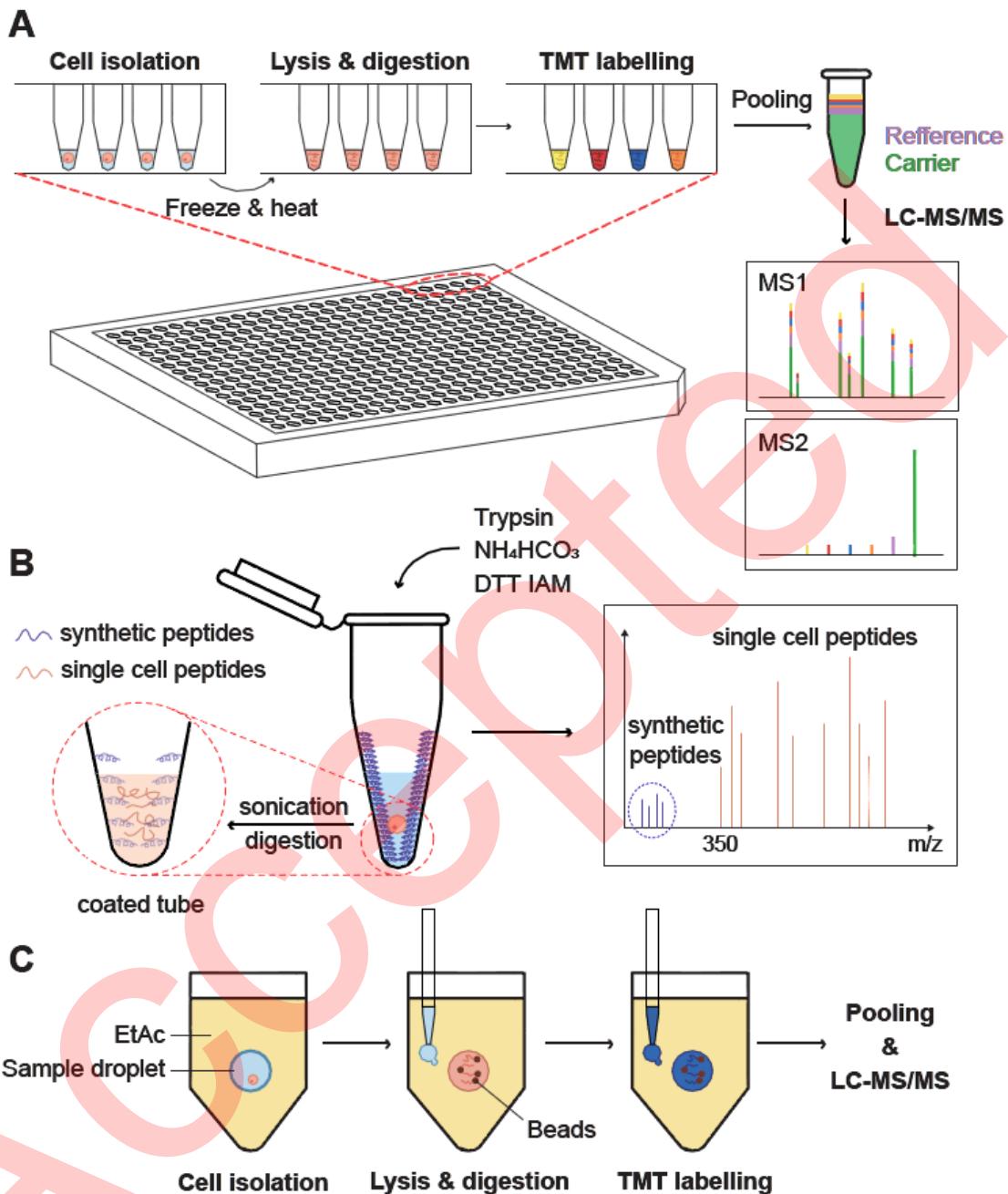


Figure 9 workflows of recent easy-to-use SCP tools. A, SCoPE2 (Specht et al., 2021). B, WinO (Masuda et al., 2022). C, Mad-CASP (Li et al., 2022g).

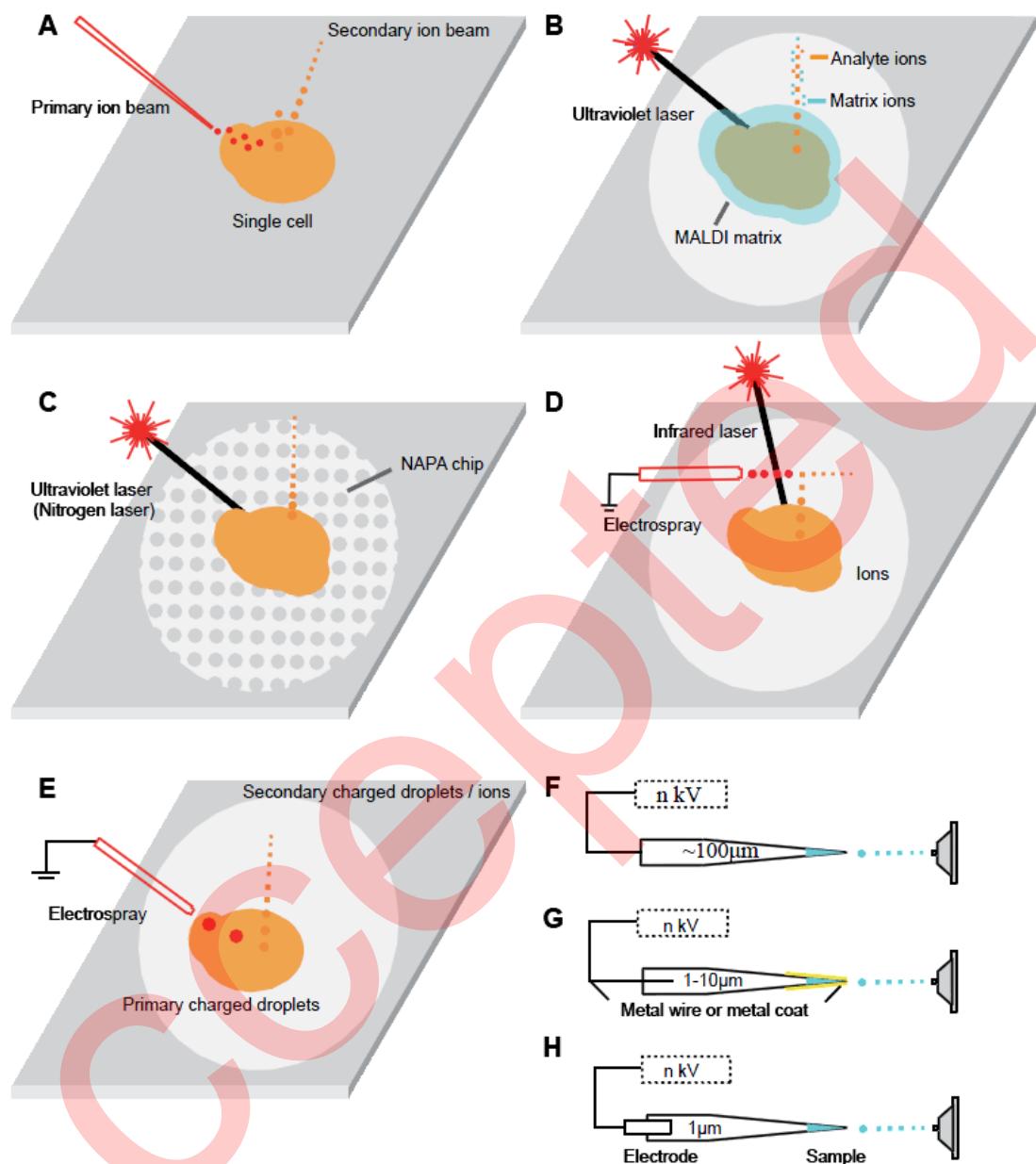


Figure 10 Examples of ionization techniques in single-cell metabolomics based on MS. A, Secondary ion MS (SIMS) (Wu et al., 2017a); B, Matrix-assisted laser desorption / ionization (MALDI) MS; C, Nanopost array (NAPA) MS (Minakshi et al., 2019); D, Laser ablation electrospray ionization (LA-ESI) (Stopka et al., 2018); E, Desorption electrospray ionization (DESI); F, Electrospray ionization (ESI); G, Nano-electrospray ionization (nano-ESI) (Bergman and Lanekoff, 2017); H, Pulsed direct current electrospray ionization MS (Pulsed-DC-ESI) (Wei et al., 2020).

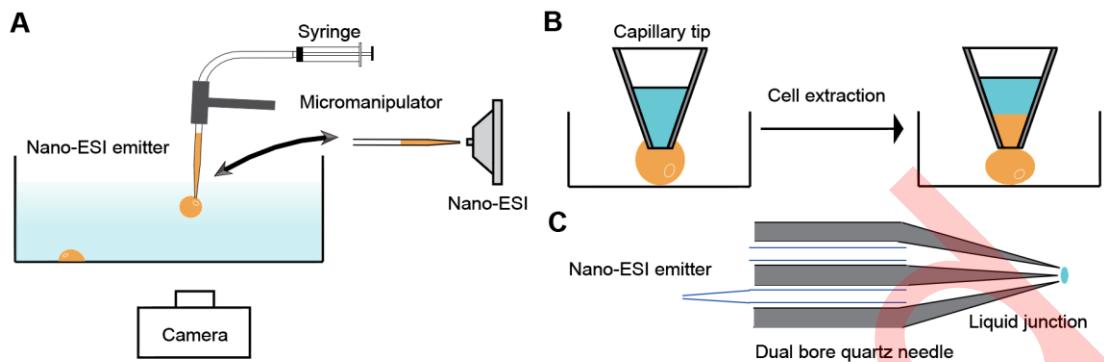


Figure 11 Examples of content extraction techniques in single-cell metabolomics based on MS. A, Live single cell MS (Live MS) (Tajik et al., 2022); B, Microextraction strategy (Yin et al., 2018); C, Single-probe diagram (Pan et al., 2020).

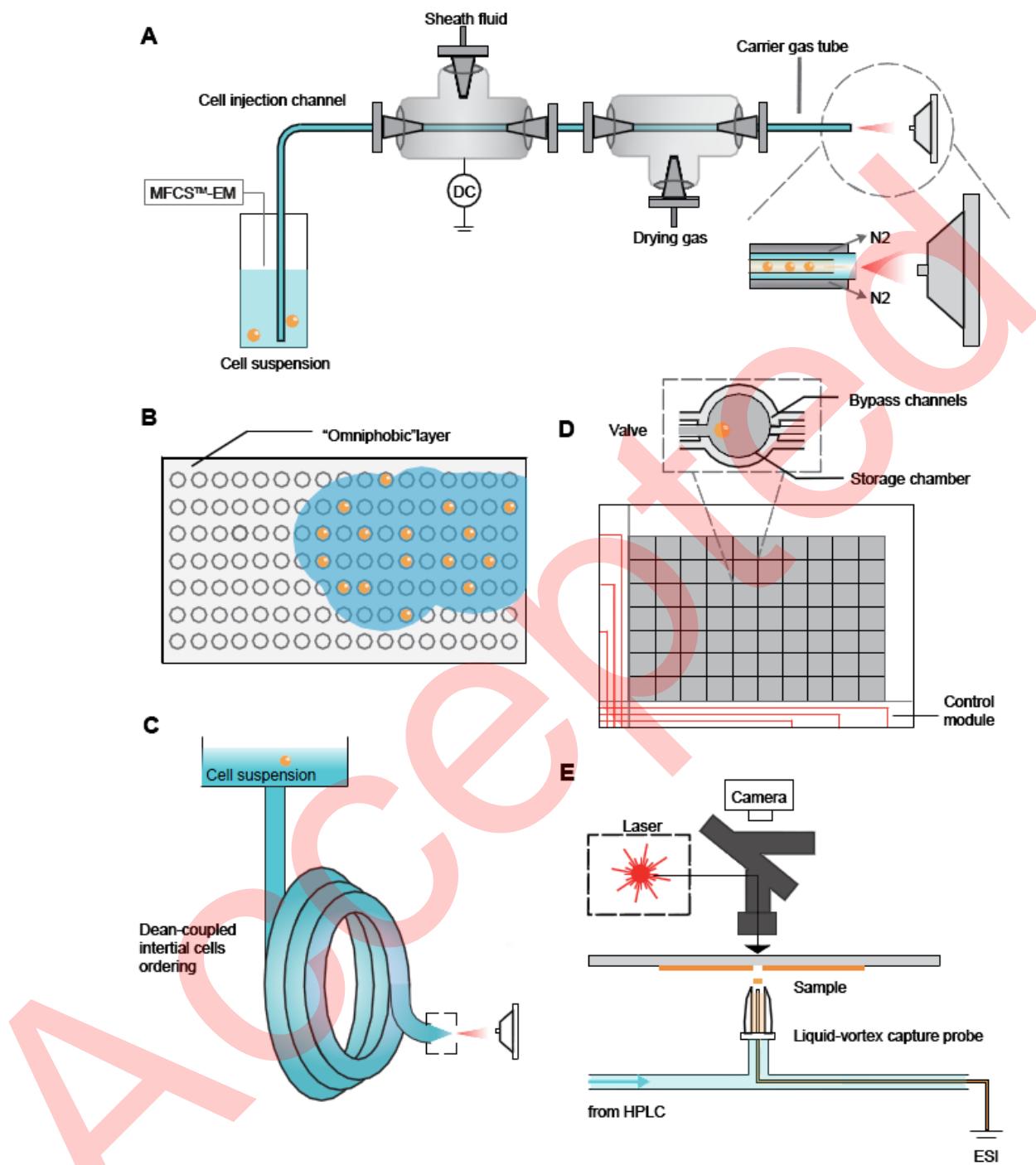


Figure 12 Examples of sorting and ionization techniques in single-cell metabolomics based on MS. A, Label free mass cytometry (CyESI-MS) (Yao et al., 2019); B, Microarrays for MS (MAMS) (Ibáñez and Svatos, 2020); C, Dean flow assisted cell ordering system (Huang et al., 2018b); D, Microvalve-based microfluidic device (Leung et al., 2012); E, Laser capture microdissection / liquid vortex capture MS (LMD-LVC-MS) (Cahill and Kertesz, 2020).

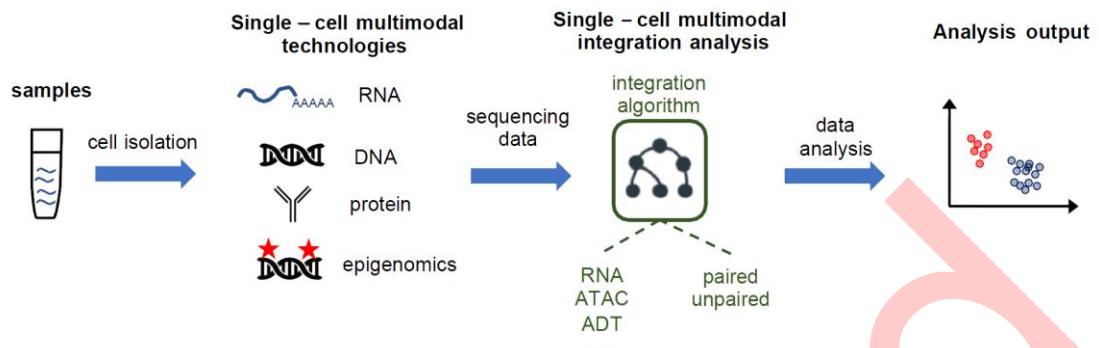


Figure 13 Major steps of single cell multimodal sequencing technology workflow.

Accepted

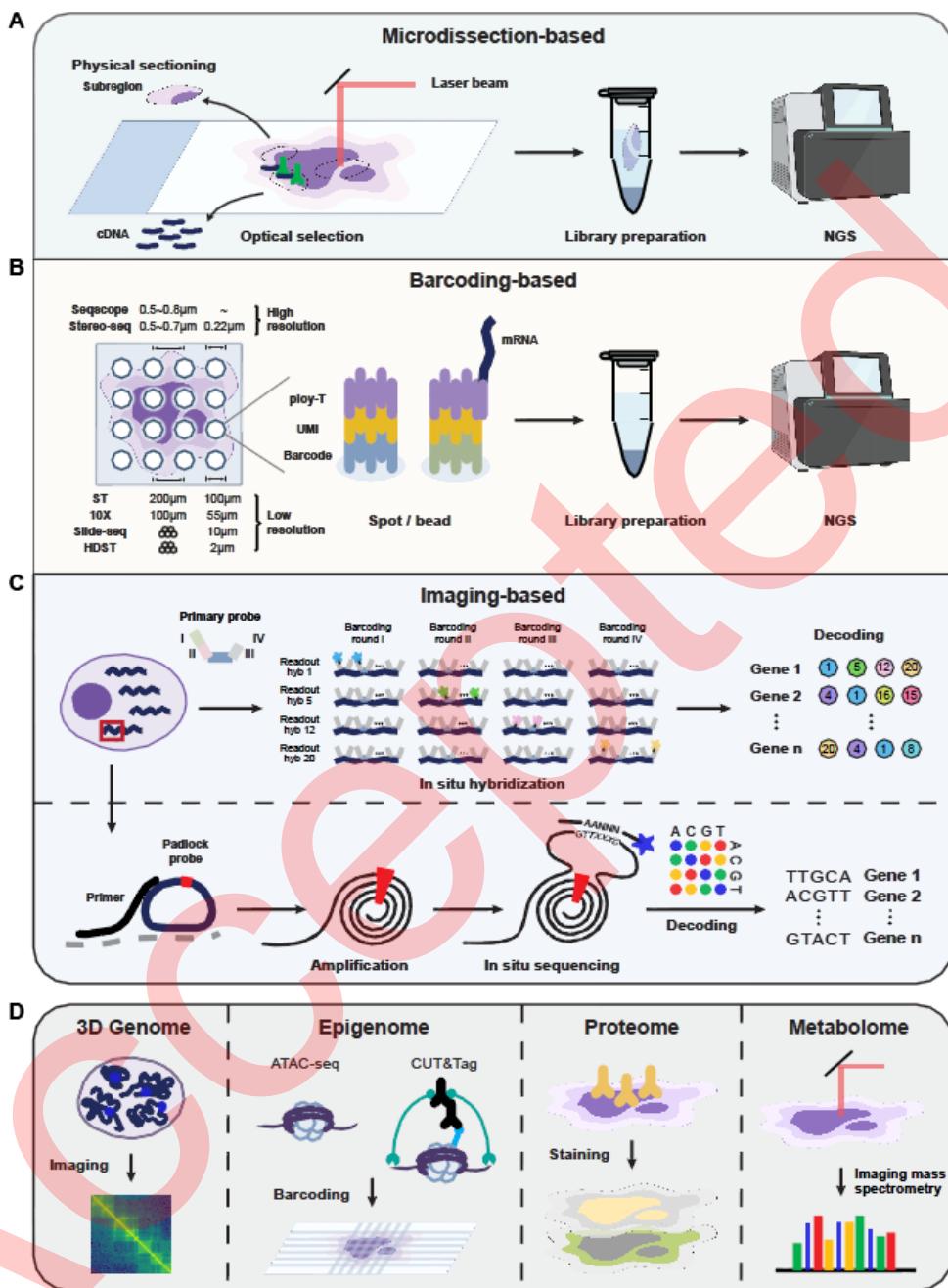


Figure 14 Schematics of techniques for spatial transcriptomics and other omics. A, For microdissection-based techniques, the sub-regions of interest can be isolated from the sample by physical sectioning or optical selection, and collected for library preparation and next-generation sequencing (NGS). B, For barcoding-based techniques, barcodes are attached to spots or beads for position labeling and *in situ* mRNA capturing. Barcoded cDNA are collected for subsequent library preparation and sequencing. C, For imaging-based techniques, with probes designed for genes of interest, *in situ* hybridization (ISH) or *in situ* sequencing (ISS) can be performed for *in situ* profiling. D, Schematics of spatial techniques for other omics, including 3D genome, epigenome, proteome and metabolome.

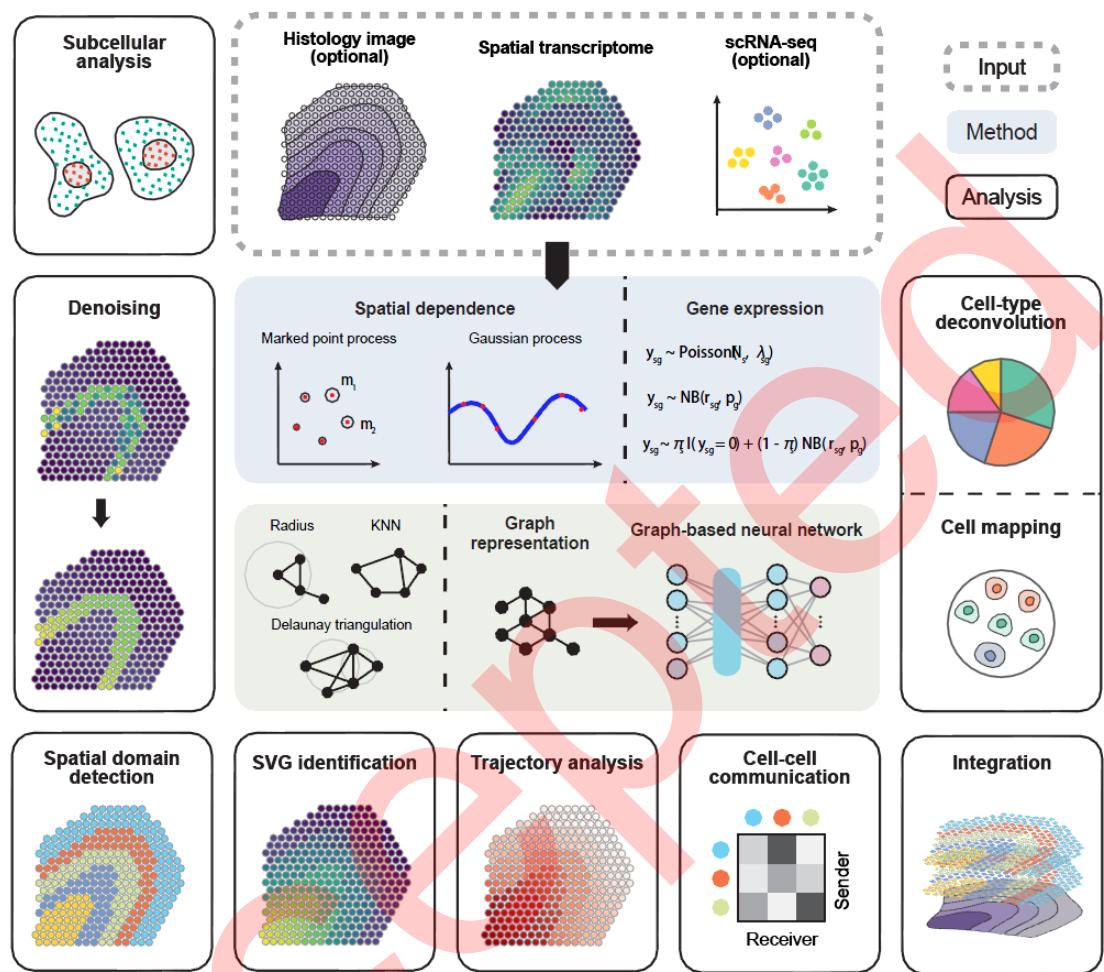


Figure 15 Overview of spatial transcriptomics analysis. With spatial gene expression taken as input, as well as optional histology images and matched scRNA-seq, a variety of analyses could be performed, including data denoising, cell type annotation, spatial domain detection, identification of spatially variable genes (SVG), pseudo-time trajectory analysis, and cell-cell communication analysis. Besides, subcellular analysis is available for high-resolution data such as imaging-based and high-resolution barcoding-based ST data. If multiple samples or multiple modalities are provided, integrative analysis can also be performed. To enable these analyses, most of the computational methods rely on probabilistic modeling or graph building to represent spatial gene expression. For probabilistic modeling-based methods, the spatial dependence between spatial locations could be modeled by marked point process or Gaussian process, and the gene expression could be modeled by Poisson, negative binomial (NB) or zero-inflated negative binomial (ZINB) distributions. For graph-based methods, neighborhood graphs could be constructed by specifying a fixed distance, or by k-nearest neighbors (KNN) or Delaunay triangulation alternatively, and then used as the input of graph-based neural networks for different analysis tasks.

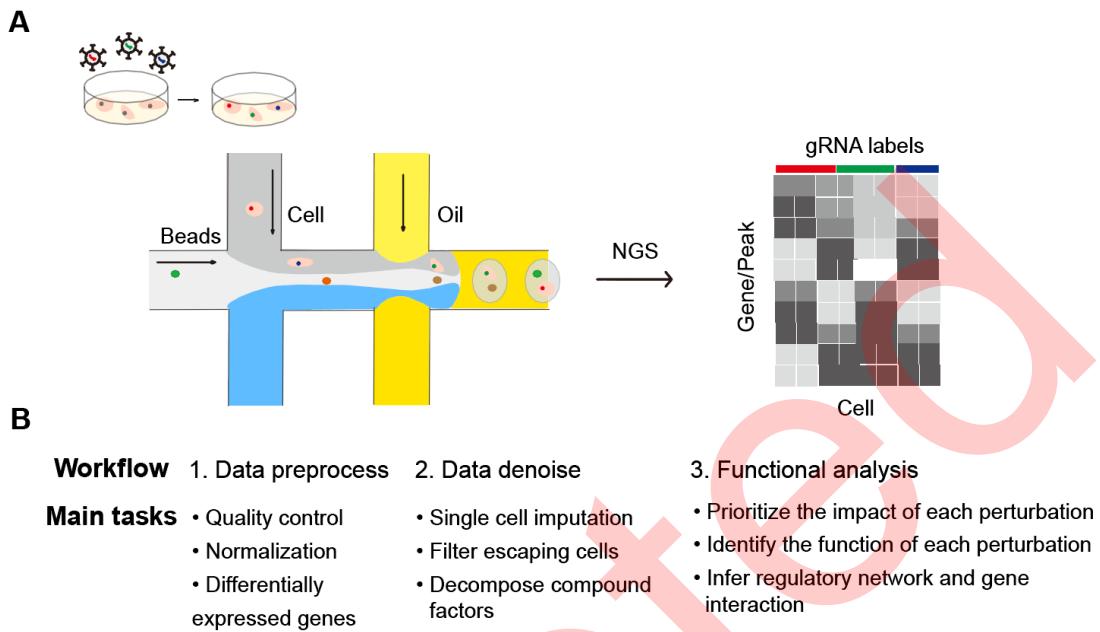


Figure 16 Overview of scCRISPR-seq. A, General schematic of scCRISPR-seq platform. NGS, next generation sequence. B, Bioinformatic analysis of scCRISPR-seq data.