BINF90008 Bioinformatics Research Project

Research Proposal

Comprehensive benchmarking of scATAC-Seq data analysis tools

Haoyu Yang (743501)

Supervisors: Matthew E. Ritchie, Shanika L. Amarasinghe

November 2019

University of Melbourne

Master of Science (Bioinformatics)

1 Introduction

Single cell sequencing technology has been widely used for understanding the heterogeneity of complex tissue and for identifying novel cell types or cell states. Previous efforts of single cell profiling are mostly performed by measuring transcriptomes using single cell RNA sequencing (scRNA-seq). scRNA-seq is relatively well developed and around 500 analysis tools are currently available for performing different tasks. In the past five years, assays for profiling the single cell chromatin accessibility landscape have emerged and provide extra information about gene regulation at the epigenetic level. Due to its simplicity and sensitivity, single cell Assays for Transposase-Accessible Chromatin using sequencing (scATAC-seq) is widely used to obtain chromatin accessibility. The timeline for scATAC-seq protocol developments and data production has been summarised by the first author of *SnapATAC* [1], as shown in Figure 1. The four main methods include the combinatorial indexing approach **(sci-ATAC-seq)** [2], microfluidics-based methods **(scATAC-seq)** [3], nano-well based protocols **(μscATAC-seq)** [4] and droplet-based **(10X scATAC-seq, dscATAC-seq and dsciATAC-seq) approaches** [5, 6]. The analysis of scATAC-seq data is somewhat different to that of scRNA-seq data and not many analysis tools are currently available for scATAC-seq data analysis. Some challenges relevant to scATAC-seq data include increased sparsity since there are only two copies of DNA that can be sequenced in diploid cells and the small amount of expected peaks [7].

The general workflow for scATAC-seq data analysis comprises (1) preprocessing: demultiplexing, adaptor trimming, read mapping, quality control, cell calling and mutiplet removal (optional); (2) feature matrix construction: defining regions via peak calling or genome binning, counting defined features, transformation and dimensionality reduction; (3) downstream analysis: cell clustering, peak calling (optional), visualisation, differential accessibility analysis and cis-regulatory network analysis [7, 8]. Several analysis methods have been generated for scATAC-seq analysis, including workflow devloped by Cusanovich *et al.* [2], *SCRAT* [9], *chromVAR* [10], *scABC* [11], *cicero*

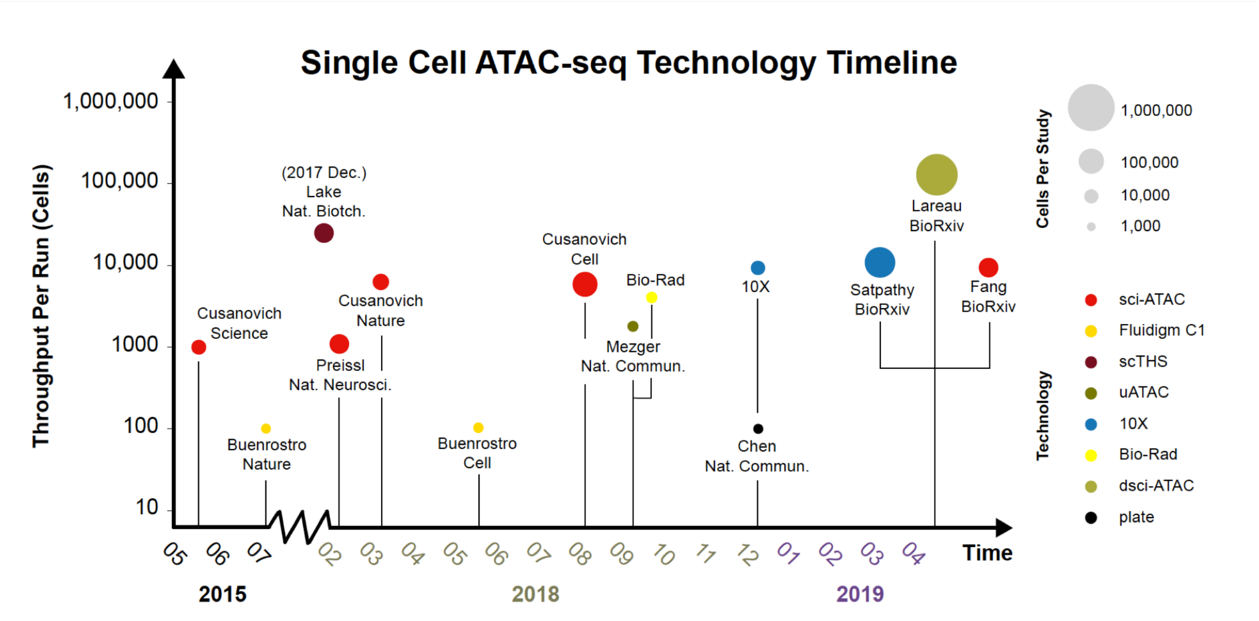

Figure 1: **Timetine of single cell ATAC-seq technology development and data generation.** Source: https://github.com/r3fang/SnapATAC/blob/master/notebooks/experiemnt_timeline.md

[12], *BROCKMAN* [13], *ScAsAT* [14], *CellRanger ATAC*, *Destin* [15], *Gene scoring* [6], *SnapATAC* [1], *cisTopic* [16], *AtacWorks* [17] and *scATAC-pro* [8].

These tools handle different steps in the analysis workflow using different algorithms. To better understand the strengths and weaknesses of different tools, it is important to benchmark their performance on data sets with known ground-truth. This allows data analysts to infer the optimal methods for each step and to develop a comprehensive pipeline for use in practice.

This project will comprehensively evaluate scATAC-seq data analysis tools and gaps in analysis workflows using in-house generated scATAC-seq data and publicly available bulk ATAC-Seq and scATAC-seq data as well as optimised universal evaluation metrics. The first part of the project comprises of adapting and developing such evaluation metrics and models for existing and potential upcoming scATAC-seq analysis tools. The second part of the project will use these validated metrics to assess several aspects of performance for existing tools, including a) usability, b) ability to handle variability in sequencing depth, c) ability to handle variations in data quality and d) the influence different starting parameters have on results. This project may further contribute to scATAC-seq pipeline development if time allows in the form of a new R-based scATAC-seq preprocessing method.

2 Literature review

2.1 Chromatin Accessibility and ATAC-seq

In eukaryotic cells, chromatin is packaged into arrays of nucleosomes, each consisting of an octamer of histones wrapped by around 147 base pairs (bp) of DNA, and separated by linker DNA [18–20]. A chromatin region is open or accessible when chromatin-binding factors can physically contact the region, which is historically characterised by nuclease hypersensitivity *in vivo* [21]. Chromatin accessibility is determined by the occupancy of chromatin by nucleosomes and the occupancy by other chromatin-binding factors, such as transcription factors (TFs), RNA polymerase and architectural proteins. In the human genome, around 2-3% of DNA sequence is accessible and more than 90% of these sequences can be bound by TFs. The post translational modification and the composition of nucleosomes change chromatin accessibility through altering the binding of TFs through steric hindrance [22] and changing the affinity of nucleosomes to chromatin re-modellers [23]. There are fewer nucleosomes at regulatory regions including enhancers, insulators and transcribed gene bodies [24, 25] so at these regions more chromatin is accessible. The positioning of nucleosomes throughout a genome affects critical cellular functions such as transcription, DNA repair and replication, as it modifies availability of binding sites to TFs, RNA polymerase and other nuclear proteins [26]. Therefore, collecting and comparing genome-wide chromatin accessibility is important for locating epigenetic changes that accompany cell differentiation, environmental signaling and disease development [27].

Chromatin accessibility is determined through quantifying the susceptibility of chromatin to either enzymatic cleavage of its constituent DNA or methylation. Some assays that directly isolate accessible regions include DNase I hypersensitive site sequencing (DNase-seq) (Figure 2a), Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) (Figure 2b), and Nucleosome Occupancy and Methylome sequencing (NOMe-seq) (Figure 2d), while the Micrococcal Nuclease sequencing (MNase-seq) (Figure 2c) indirectly measures chromatin accessibility. **DNase-seq** uses the endonuclease DNase to cleave DNA within accessible chromatin (Figure 2). Boyle *et al.* (2008) [29] used a type II restriction enzyme to make a single cut and then ligate adaptors, whereas Hesselberth *et al.* (2009) [30] used enzymes to make double cuts and applied size selection. **ATAC-seq** uses hyperactive transposases (Tn5) to simultaneously cleave and ligate adaptors to accessible DNA. Cleaved fragments with two different adaptors will be PCR amplified and after size selection, short fragments will be sequenced [31]. **MNase-seq** uses the endo-exonuclease MNase to both cleave and eliminate accessible DNA. Thus, regions occupied by nucleosomes and other chromatin-binding factors are isolated and sequenced [32, 33]. **NOMe-seq** uses GpC methyltransferase to methylate accessible DNA. Non-methylated cytosines in the sequence will be converted to uracil following bisulfite conversion, and only accessible regions are sequenced [34].

2.2 single-cell ATAC-seq (scATAC-seq)

The rapid development of protocols for single-cell RNA-seq profiling has greatly increased our understanding of existing cells and also led to the discovery of new cell types. However, an important piece of information that is often missing from single-cell RNA-seq is the cell-specific genomic regulatory model. Such regulatory model comprises of heterogeneous enhancers, promoters, and insulators that are important for the modulation of single-cell gene expression in a spatiotemporal continuum. Due to its simplicity and sensitivity, in recent years, ATAC-seq has been widely used to measure single-cell chromatin accessibility.

scATAC-seq data can be generated by four major approaches: a) combinatorial indexing through split-pooling [6, 35], b) microfluidics-based methods [3], c) nano-well-based methods [4] and d) droplet microfluidics [5, 6] (Figure 3). In the combinatorial indexing method **(sci-ATAC-seq)** (Figure 3a), the isolated nuclei are split into wells and uniquely barcoded with Tn5 transposases for the first round. These nuclei are pooled and redistributed randomly into wells for the second round of barcoding. This method takes advantage of the low probability of having two cells in

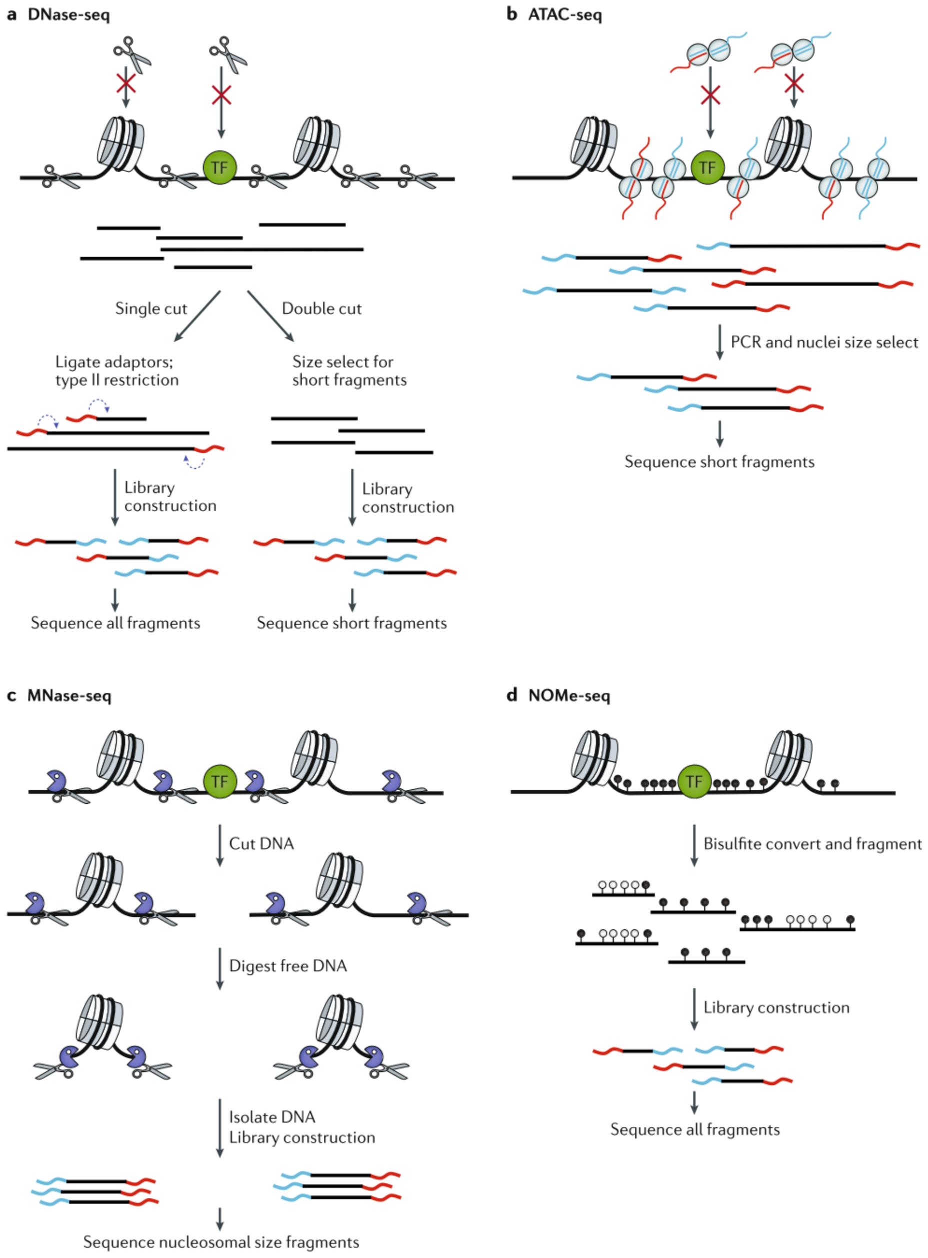

Figure 2: **Principal methods of measuring chromatin accessibility, adapted from Klemm, Shipony and Greenleaf (2019)** [28]. (a) DNase-seq uses endonuclease DNase to cleave accessible DNA regions (indicated by scissors on the DNA). At the protein-bound position, chromatin are protected against endonuclease cleavage (indicated by the red crosses). Following cleavage, the DNA fragments can be either ligated with adapters at one end after the digestion of type II restriction enzyme (single cut) or digested at both ends and selected for short fragments (double cut). (b) ATAC-seq takes advantage of Tn5 transposases which cleave and ligate adaptors to accessible DNA. (c) In MNase-seq, the MNases, which are both endonuclease and exonuclease, cleave and digest accessible DNA and the inaccessible DNA fragments are sequenced. (d) In NOMe-seq, accessible DNA are methylated (indicated by the black pins) by GpC methyltransferase followed by bisulfite conversion of non-methylated cytosine to uracil. Therefore, the presence of cytosine indicates that the corresponding DNA regions are accessible.

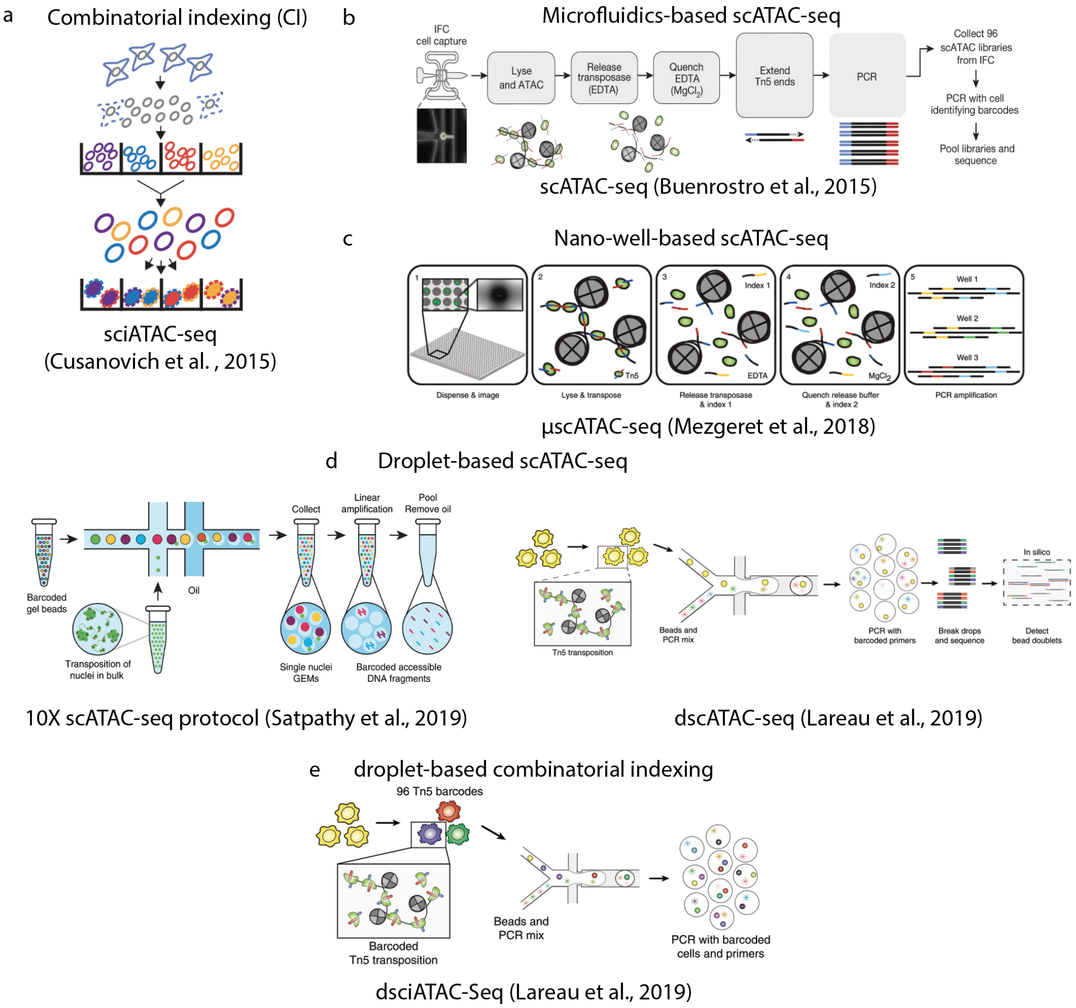

Figure 3: **Protocols for single-cell ATAC-seq (scATAC-seq)** [3–6, 35, 36]. (a) In combinatorial indexing methods, nuclei are isolated and molecularly tagged in bulk with barcoded Tn5 transposases in wells (different barcodes are represented by the different colors outlining the nuclei). Nuclei are then pooled and redistributed into wells and a second barcode (represented by the color filling each nucleus) is introduced during PCR (b) In microfluidics-based scATAC-seq, individual cells are captured using microfluidics platform (Fluidigm). The isolated cells are transposed, and fragments are amplified by PCR on the integrated fluidics circuit (IFC). Then in 96-well plates, with each well containing one scATAC library, the scATAC libraries are amplified and barcoded through PCR. (c) In nano-well-based scATAC-seq (µscATAC-Seq), individual cells are isolated in nano-wells and the cell-specific transposition barcoding and PCR amplification is carried out to build scATAC libraries within wells (d) In droplet-based 10X scATAC-seq and dscATAC-Seq, a pool of transposed nuclei is loaded into a droplet-microfluidics system and they are simultaneously encapsulated into single droplet emulsion with PCR reagents and barcoded gel beads. Following droplet PCR, droplet-specific barcodes are added to transposed DNA and the scATAC libraries are generated (e) In dsciATAC-seq, nuclei are distributed in wells similar to the combinatorial indexing method and transposed with well specific barcodes. The transposed nuclei are then pooled and loaded in droplet microfluidics device and the rest of processes are the same as dscATAC-seq.

the same wells in both rounds, thus there is no need for isolation of cells [35]. This method has been successfully used to study the embryonic development in *Drosophila melanogaster* [37], to study transcriptional regulation of developing mouse forebrains [38] and create a single-cell atlas across 13 adult mouse tissues [2]. **scATAC-seq**, developed by Buenrostro *et al.* (2015) enables the capture of single-cell nuclei through a microfluidic device (Fluidigm, C1) [3] (Figure 3. This method has been used to profile thousands of cells during hematopoiesis [36, 39]. The nano-well-based **µscATAC-seq** (Figure 3c), further increases the throughput of measurements [4]. Individual cells are isolated in nano-wells. In each well, cell-specific transposition barcoding and PCR amplification to build scATAC libraries are performed. Even though the scalability is poor for both scATAC-seq approaches compared to the combinatorial indexing method, the single-cell library complexity is higher, which is critical for reducing the sparsity of scATAC-seq results [3, 28, 35, 36].

Bio-Rad Laboratories [6] (Figure 3d) developed droplet-based microfluidic method, which provide similar data quality compared to the previous two microfluidic methods [28]. It also provides an option that allows multiple gel beads captured in the same emulsion drop to increase throughput and a demultiplexing analysis tool was developed to address the problem of one cell having multiple barcodes [6]. To further increase throughput, combinatorial indexing and droplet methods were used together in the **droplet-based single-cell combinatorial indexing for ATAC-seq (dsciATAC-seq)** [6] approach (Figure 3e).

The protocol developed by **10X Genomics** (i.e. 10X-ATAC) [5] is in general similar Bio-Rad dscATAC-seq method. In the 10X method, nuclei are first isolated and transposed with Tn5 transposase. Then each nucleus will be encapsulated by a droplet with a gel bead containing barcodes. After linear amplification, the DNA fragments are barcoded and later emulsion break open the droplets to allow the barcoded DNA to be pooled for PCR amplification and high-throughput sequencing [5]. The main difference between 10X-ATAC and dscATAC-seq is that dscATAC-Seq allows multiple beads in one droplet to increase library complexity, while 10X-ATAC provides larger close-packed hydrogel beads which allow a certain level of control over the amount of beads loaded into one droplet.

2.3 Characteristics of scATAC-seq data

scATAC-seq data is highly sparse compared to scRNA-seq data. For an expressed gene, there may be several copies of RNA molecules within a cell to be sequenced while there are only a few copies of DNA (two in diploid organisms) for scATAC-seq assays. This results in the detection of only 1 - 10% of the expected accessible peaks in the Fludigm C1 platform [3]. Therefore, the method used to recover informative features from such sparse data is critical for measuring chromatin accessibility using this technology [7].

The library structure generated by 10X-ATAC [5], dscATAC-seq [6] and sciATAC-seq [35] are shown in Figure 4. The 10X-ATAC protocol generates four `fastq` files which are read 1 (50 bp), read 2 (50 bp, reverse direction), the 10x cell barcode (16 bp) and sample index (8 bp). The library structure generated by dscATAC-seq [6] is similar to 10X scATAC library in general, that is the cell barcode is at one side of the chromatin fragment and the sample index is on the other side. But the cell barcode is composed of three 7 bp barcode fragments. One `fastq` file is 118 bp containing cell barcode, read 1 and other sequences. The other two `fastq` files are 40 bp read 2 in reverse direction and 8bp sample index. The library structure of sciATAC-seq [35] is different, with the two fragments of barcod on both side of chromatin fragment. Two of the `fastq` files are read 1 and read 2 in reverse direction. One `fastq` file contains half of barcode and other sequences. The rest `fastq` file contains the other half of barcode and sample index.

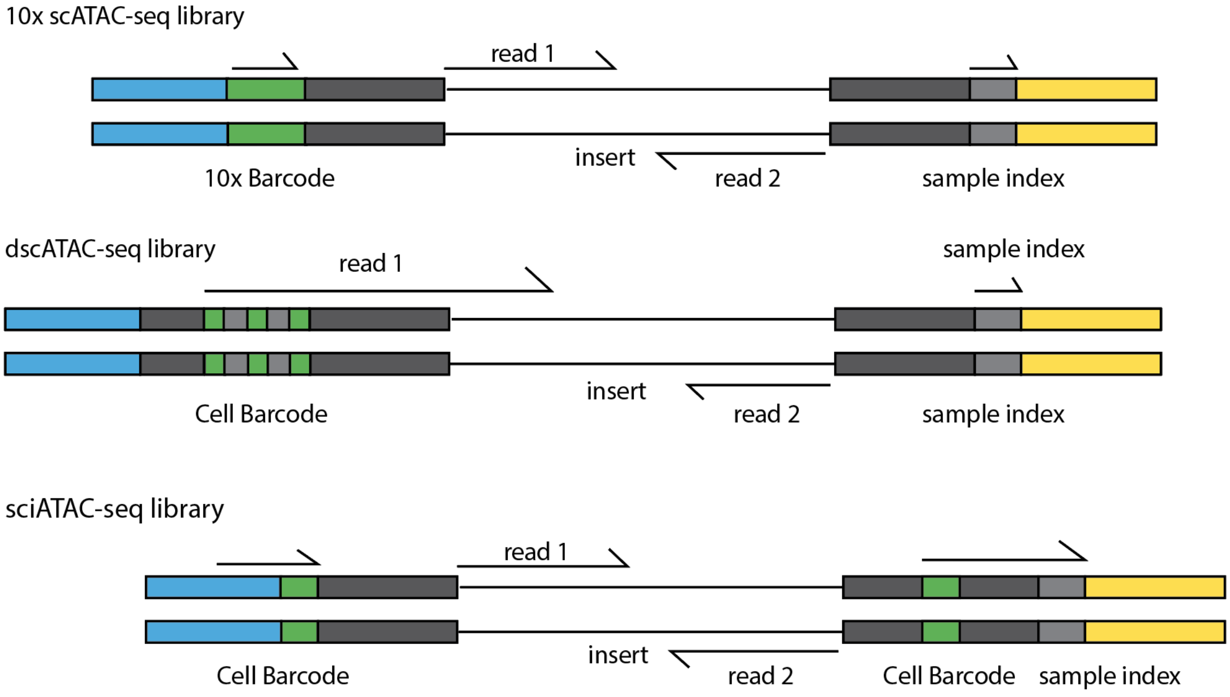

Figure 4: **The expected library and read structure generated from 10X-ATAC [5], dscATAC-seq [6] and sciATAC-seq [35].** The green blocks are cell barcodes, the light gray blocks are sample indexes and arrows indicate reads generated by sequencing.

2.4 Example applications of scATAC-Seq

Single cell ATAC-seq has been applied to answer a range of different biological questions. Here I summarise a number of recently published data sets generated from scATAC-seq applications and describe the major findings from each study.

2.4.1 scATAC-seq of human hematopoietic cells (Buenrostro *et al.*, 2018)

The data setwas generated on 2,034 cryopreserved CD34$^+$ cells in bone marrow from 6 human donors. The cells were first isolated by fluorescence activated cell sorting (FACS) followed by cryopreservation. Then scATAC-seq [3] was performed on each cell type. The cells comprise of hematopoietic stem cells (HSCs), multipotent progenitors (MPPs), lymphoid-primed multipotent progenitors (LMPPs), common-myeloid progenitors (CMPs), granulocyte-macrophage progenitors (GMPs), megakaryocyte-erythrocyte progenitors (MEPs), common-lymphoid progenitor (CLPs), plasmacytoid dendritic cells (pDCs), monocytes, and an uncharacterised CD34$^+$CD38$^-$CD45RA$^+$CD123$^-$ population. The median number of fragments per cell was 8,268 with 76% of them mapping to peaks, resulting in a median of 6,442 fragments in peaks per cell [39].

Buenrostro *et al.* (2018) constructed a chromatin accessibility landscape of human hematopoiesis to characterise

Table 1: **Examples of publicly available scATAC-seq data sets**

Data set	Methods	Properties	Cell Numbers	Reference
Human hematopoietic lineage	scATAC-seq	continuous	2,034	Buenrostro *et al.*, 2018, GSE96772
Drosophila **embryos**	sciATAC-seq	continuous	23,085	Cusanovich *et al.*, 2018, GSE101581
Mouse Cell Atlas	sciATAC-seq	large data set	81,173	Cusanovich *et al.*, 2018, GSE111586
Human hematopoeitic cells	10X-ATAC	large data set	61,806	Satpathy *et al.*, 2019, GSE129785, 10X-ATAC PBMC
Basal cell carcinoma tumor microenvironment (TME)	10X-ATAC	cancer	37,818	Satpathy *et al.*, 2019, GSE129785

differentiation trajectories. They use ChromVAR to infer TF activity by calculating TF motif-associated chromatin accessibility changes of each cell. Based on this analysis, they found that HSCs exhibit low levels of lineage specifying motifs and high levels of motif regulating stem cell activity, but these were reversed in more differentiated cells. They also observe heterogeneity within CMPs and GMPs and develop a strategy to partition GMPs along their differentiation trajectory. Furthermore, they integrated scRNA-seq data with scATAC-seq date to associate transcription factors to chromatin accessibility changes through correlations of expression and regulatory element accessibility, which provided a computational method for integrative exploration of complex regulatory dynamics in a primary human tissue at single-cell resolution [39].

2.4.2 sciATAC-seq of *Drosophila* embryos (Cusanovich *et al.*, 2018)

Using combinatorial indexing assay (sciATAC-seq), Cusanovich *et al.* (2018) profiled chromatin accessibility in 23,085 single nuclei from hundreds of fixed *Drosophila melanogaster* embryos across three landmark embryonic stages: 2 to 4 hours after egg laying (AEL), 6 to 8 hours AEL, and 10 to 12 hours AEL. They revealed the spatial heterogeneity in chromatin accessibility of regulatory genomic regions before gastrulation, which aligns with the future cell fate. During mid embryogenesis, cell types can be inferred by their chromatin accessibility, while maintaining a signature of their germ layer of origin. They identified over 30,075 distal elements with tissue-specific accessibility. Their work demonstrated the power of scATAC-seq in profiling of embryos to resolve dynamic changes during development, and to uncover the cis-regulatory programs of germ layers and cell types [37].

2.4.3 sciATAC-seq atlas of mouse tissues (Cusanovich *et al.*, 2018)

Cusanovich *et al.* (2018) applied the combinatorial indexing assay, sciATAC-seq, to profile genome-wide chromatin accessibility in 81,173 single cells from adult mouse tissues. The data set was generated on 13 tissues from 5 8-week-old male C57BL/6J mice, including bone marrow, cerebellum, heart, kidney, large intestine, liver, lung, prefrontal cortex, small intestine, spleen, testes, thymus, and whole brain, using sciATAC-seq methods [35]. For tissues from bone marrow, large intestine, lung and whole brain, a replicate sample from a second mouse were collected. The total number of cells profiled per tissue (after filtering) ranged from 2,278 for cerebellum to 9,996 for lung (two samples).

They identified and annotated 85 distinct patterns of chromatin accessibility based on the accessibility score of each cell at the predefined 436,206 potential regulatory elements using t-distributed Stochastic Neighbor Embedding (t-SNE) and Louvain clustering. Besides this, they linked regulatory elements to their target genes using Cicero [12] to define the TF motif specifying each cell type and to identify heterogeneity within cell types. Furthermore, they developed a strategy for mapping scRNA-seq data to sciATAC-seq data, in order to facilitate the comparison of atlases. Finally, they identified cell-type-specific enrichments of the heritability signal for hundreds of complex traits, by intergrating mouse chromatin accessibility with human genome-wide association summary statistics [2].

2.4.4 Human hematopoeitic cell and basal cell carcinoma tumor microenvironment (TME) study (Satpathy *et al.*, 2019)

Satpathy *et al.* (2019) generated scATAC-seq profiles of 61,806 cells from peripheral blood and bone marrow from 16 healthy individuals using the commercial system 10X scATAC-seq . The cell types range from bone marrow progenitor cells to multiple types of differentiated immune cells including B cells, T cells and NK cell. They identified 31 clusters by performing Latent Semantic Indexing (LSI) and Shared Nearest Neighbor (SNN) clustering [40]. Then in order to classify these clusters, they performed three strategies: (1) classified clusters based on the neigbouring genes of cluster-specific *cis*-elements; (2) calculated gene activity scores, which are the aggregate accessibility of several enhancers linked to a single gene promoter [12] and (3) used TF motifs, computed from the accessibility of TF binding sites genome-wide in each single-cell [10]. They also demonstrated that using the first strategy can identify cell type-specific cis-elements even for a single gene. All of these strategies did not involve pre-labelling of cell types [5].

In basal cell carcinoma, they generated scATAC-seq profiles of 37,818 cells from biopsies of pre- and post- anti-programmed cell death protein 1 (PD-1) treatment from 7 patients for cell types including T cells, non-T immune cells, stromal cells and tumor cells. After analysis, they revealed the regulatory elements in cancer, stromal and immune cells in the tumor microenvironment. By comparing cells from before and after PD-1 treatment, they identified cell types that were sensitive to the therapy and revealed a shared regulatory program that governs intratumoral CD8+ T cell exhaustion and CD4+ T follicular helper cell development [5].

2.5 A general workflow for scATAC-Seq data analysis

The general workflow for scATAC-seq data processing includes three major steps: (1) data pre-processing, (2) feature matrix construction and (3) downstream analysis (Figure 5).

2.5.1 Data pre-processing

After sequencing, the `BCL` files (Illumina's sequencer basecalled file format) need to be converted into `fastq` format. One common tool for this is *bcl2fastq*. The next step is demultiplexing, where the barcode from the barcode `fastq` file is added to the name section of each read to facilitate further analysis. Next the `fastq` files are generally passed through a general quality control (QC) step and adapter trimming (optional). Lots of tools have been developed for

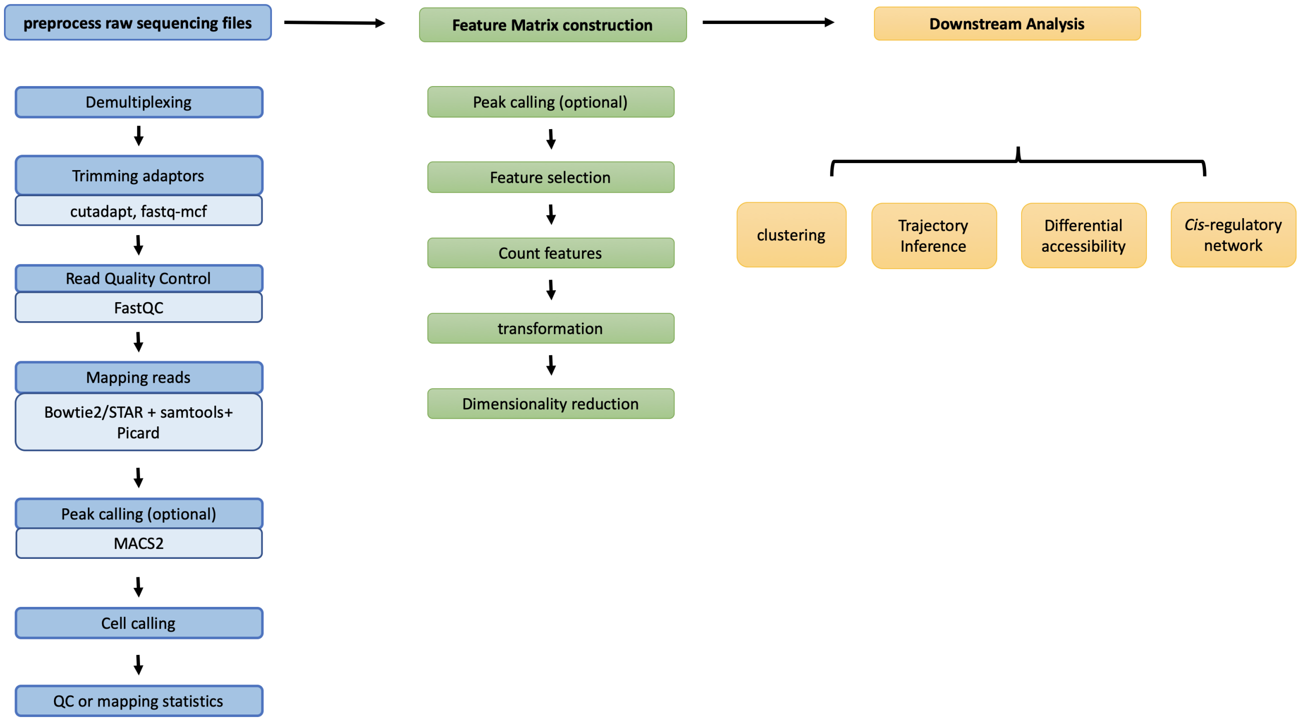

Figure 5: **A general workflow for scATAC-seq data analysis, adapted from Chen *et al.* (2019) [7]**

this step such as *FastQC* [41], *cutadapt* [42] and *Trim Galore!* [43]. The remaining fragments are then aligned to a reference genome using programs such as *BWA* [44] and *bowtie2* [45]. After mapping, *samtools* [46] can be used to filter out low quality reads, sort, index and mark duplicates.

Mapping statistics and quality control metrics can then be calculated. Commonly calculated metrics include the total number of reads, the total number of mapped reads, the mapping rate, fraction of reads mapping to the mitochondrial genome, number of duplicate reads, number of high-quality reads, library complexity, fraction of reads in annotated genomic regions and the TSS enrichment profile [8].

Cell calling is the process of identifying barcodes that represent 'real' cells versus those that represent noise. This step is essential because not all barcodes represent true cells due to cell collisions and/or cell debris. This is still a challenging step and different pipeline or analysis strategies use different methods. The first method is based on using summary statistics such as the total number of fragments per barcode or fraction of barcodes in peak regions, thus low quality barcodes will be filtered out [8]. The second method is a model-based approach adapted by *Cell Ranger ATAC* [47], which fits a mixture model of two negative-binomial distributions to distinguish cell barcodes from non-cell barcodes. After cell calling, additional QC can be performed using previously described statistics, which can include the number of cells called, the median number of fragments per cell and the percentage of mapped reads with cell barcodes [8].

2.5.2 Peak calling

Peak calling can occur as part of feature matrix construction or during downstream analysis. It is a process that identifies regions that are enriched by the reads or fragments. *MACS2* [48] is a commonly used peak calling method that has been adapted by many analysis tools or workflows. *ZINBA* [49] is another method that is adapted by *Cell Ranger ATAC* [47]. The output of this step is the peak by barcode matrix [8].

2.5.3 Feature matrix construction

Generating an accurate feature matrix is crucial for correct downstream analyses of scATAC-seq data. Multiple approaches developed so far to achieve this step are focused on either, binning of reads to regions of the genome [1] or peak calling and binning the peaks to regulatory regions [10]. As summarised in Chen *et al.* (2019) [7] the process of constructing a feature matrix includes 1) defining regions, 2) counting features, 3) transformation and 4) dimension reduction. Different methods are developed for each sub-step and the current tools combine them in different ways (Figure 5).

Defining regions

There are 5 main types of regions defined by current scATAC-Seq tools; a) peak regions called on respective bulk ATAC-Seq data [10–12, 14, 16], b) peak regions on aggregated single-cells [10–12, 14, 16], c) transposon integration sites (*BROCKMAN* [13]) d) known chromatin regions containing certain TF motifs (*SCRAT* [9]) e) whole genome sectioned into uniform bins (*SnapATAC* [1]) or a subset of it [2]). It has to be noted that defining regions using known regions and/or peaks identified from true bulk or pseudo-bulk regions may fail to call peaks that are only observed in rare cell types.

Counting features

Raw feature counting within the defined regions can be achieved by either by a) counting the reads overlapping the regions (i.e. peaks, bins, TF motifs), b) counting k-mers or gapped k-mers overlapping the regions [10, 13]. Availability of a gene annotation is used by some tools to generate a gene enrichment score (i.e. gene activity) based on reads nearby to a specific gene [6, 9, 12].

Transformation

Most methods convert the feature matrix into a binary matrix assuming that each feature could be either ”open” or ”closed” in accessibility (for diploid organisms) and the resulting count matrix is extremely sparse [1, 2, 12, 14, 16]. An advantage of a binary matrix is it allows one to overlook the technical issues arising from low sequencing coverage and PCR artifacts. Transformations that are currently enforced on these binary matrices by existing tools include the Jaccard index [1, 14], term frequency inverse document frequency transformation (TF-IDF) [2], weighting by co-accessibility [12], weighting by a decaying function based on the distance to gene TSS [6], sample depth correction [1, 10], z-scores to measure the gain and loss of chromatin accessibility across cells [10, 13].

Dimensionality reduction

Principal component analysis (PCA) is the most widely used dimensionality reduction method [1, 2, 13] while some studies use latent Dirichlet allocation (LDA) to construct features to select for dimensionality reduction and others use multidimensional scaling (MDS) [14].

2.5.4 Downstream analysis

Finally, downstream analysis can include clustering, visualisation, peak calling (optional), trajectory inference, differential accessibility analysis and cis-regulatory network construction.

The goal of clustering is to classify cells into sub-groups based on the feature matrix obtained in the previous step. Common methods include Model-Based Clustering (mclust) [50], DBSCAN, Hierarchical Clustering, K-means, weighted K-medoids [51], Louvain [52] etc.

In some tools, peak calling is performed at this stage to reveal peaks in new cell types or in rare cell types.

For visualisation, high dimensional data are projected onto a 2D plane. t-Distributed Stochastic Neighbor Embedding (tSNE) [53] and Uniform Manifold Approximation and Projection (UMAP) [54] are commonly used in most of analysis tools for this task.

Trajectory inference is to infer the cell differentiation trajectory based on the accessibility changes of the cells. Differential accessibility analysis aims to identify cell-type specific regulatory elements. *Cis*-regulatory network analysis identifies annotated and potential enhancer regions of certain gene promoters, which is also known as co-accessibility analysis.

2.6 Existing scATAC-seq data analysis tools

Table 2: **Currently available single-cell ATAC-seq analysis tools**

Method	**References**	**Availability**
Latent Semantic Indexing; LSI	Cusanovich *et al.*, 2015, 2018	-
chromVAR	Schep *et al.*, 2017	R package
SCRAT	Ji *et al.*, 2017	GUI, R package
scABC	Zamanighomi *et al.*, 2018	R package
Cicero	Pliner *et al.*, 2018	R package
BROCKMAN	de Boer & Regen, 2018	R package
Scasat	Baker *et al.*, 2018	R package
Cell Ranger ATAC	10X Genomics, 2018	-
Destin	Urrutia *et al.*, 2019	R package
Gene Scoring	Lareau *et al.*, 2019	workflow combines python and R
SnapATAC	Fang *et al.*, 2019	Python, R package
cisTopic	Bravo *et al.*, 2019	R package
SCALE	Xiong *et al.* 2019	Python
AtacWorks	Lal *et al.* 2019	linux
scATAC-pro	Yu *et al.* 2019	Python, R, linux

Cusanovich *et al.* [**2**] generated a analysis workflow that has been used to analyse a mouse cell atlas data set and a *drosophila* embryonic development data set. Before peak calling, it clusters cells using latent semantic indexing (LSI). A bin by cell binary matrix is first constructed as mentioned in Section 2.4 (2). Then commonly used bins are selected and cells with low read counts are filtered out. Normalisation of reads is performed using the term frequency-inverse document frequency transformation (TF-IDF) followed by dimensionality reduction using singular value decomposition (SVD). After these steps, the first-round of clustering is performed (referred to as *'in silico* *cell sorting'*) to generate clades. To identify specific regulatory elements within each clade, peaks are called using *MACS2* within each cluster, and combined into one .bed file. Finally, the clusters are refined with a second-round of clustering after TF-IDF and SVD based on read counts from the peaks called previously.

SCRAT [9] was designed for analysing single-cell regulome data, with an online web graphical user interface and an R package. It covers the preprocessing, feature summarisation, clustering, differential accessibility analysis and infer cell identity. The input are `bam` files and an option is provided to exclude any signal from the ENCODE blacklist regions [55]. It combines read counts on different regulatory features such as TF binding motifs, gene TSS regions and user defined features. It provides PCA and t-SNE as dimension reduction methods and multiple clustering methods including Model-Based Clustering (mclust) [50], DBSCAN, Hierarchical Clustering and k-means. It can also perform differential accessibility analysis. For each feature, statistical tests including parametric (t-, ANOVA F-) or non-parametric (Wilcoxon rank-sum, Kruskal-Wallis or permutation) tests can be performed between the clusters, to identify differential features. Features that pass a particular false discovery rate threshold are then reported.

chromVAR [10] takes input of aligned sequences, a peak file (determined from either bulk reference or aggregated single-cell data) and then estimates the dispersion of chromatin accessibility within peaks sharing the same feature, e.g. TF motifs or k-mers. It sampling 'background' peaks from all peaks for each defined feature, and the background peaks have the same GC content and fragment count as the observed peaks. The 'background' peaks are used to compute bias-corrected accessibility deviation or z-score among all cells for each defined feature. This accessibility deviation or z-score is used for downstream clustering. For downstream analysis, firstly it has an interactive web application that can show clusters and deviation score based coloring in all clusters for selected gene. Secondly, it can analyse the correlation and potential cooperativity between TF binding sites. Thirdly, it can sort features based on their variability across cells.

scABC [11] takes `bam` file and peak file obtained from aggregation of all cells. It first calculates a global weight for each cell based on the number of distinct reads in the peak regions. Based on these weights, it then uses weighted k-medoids [51] to cluster cells based on the read counts within peaks. Then to improve clustering, it calculates landmarks for each cluster, which are the P peaks with highest read count, where P is user defined. The assignment of cells to sets of landmarks (i.e. clusters) is based on Spearman correlation. Further, differential accessibility is obtained using an empirical Bayes regression based hypothesis testing procedure.

Cicero [12] also first groups similar cells by calculating a gene activity score based on accessibility at a promoter region and the regulatory potential of peaks nearby. The special feature of **Cicero** is that it identifies all co-accessible regions in order to build a *cis*-regulatory map.

BROCKMAN [13] represents genomic sequences by gapped k-mers within transposon integration sites and infers the variation in k-mer occupancy using principal component analysis (PCA). It is designed for identifying differentially active TFs and TF-TF interactions.

Cell Ranger ATAC is a set of analysis pipelines for Chromium scATAC-seq data developed by 10X Genomics. It also uses peak information for clustering with a theoretical disadvantage of not being able to identify rare peaks appearing only in very rare cell populations. It supports multiple dimension reduction methods, including PCA, LSA and PLSA and multiple clustering methods, such as k-means, graph-clustering for PCA and Spherical k-means, graph-clustering for LSA and PLSA.

Destin [15] first generates two weight peak matrices. The first one up-weights the distal regulatory regions over proximal regulatory regions based on their distances to the TSS. The second up-weights the less shared accessibility peaks with reference to chromatin accessibility peaks using DNase I hypersensitive site (DHS) data [55]. The two matrices are multiplied and weighted PCA is performed followed by k-means clustering to group cells with similar chromatin accessibility profiles.

The Gene Scoring method [6] assigns each gene an accessibility score by summarising peaks near its TSS and weighting them by an exponential decay function based on their distances to the TSS.

SnapATAC [1] is a pipeline for analysing scATAC-seq data. Before peak calling, it constructs a cell by bin matrix, coverts it into a Jaccard index matrix by measuring similarities between cells and adjusts for differences in library size using a regression-based normalisation method. The normalised matrix is used for clustering and peak calling is performed on each cluster to identify *cis*-regulatory elements.

cisTopic [16] uses Latent Dirichlet Allocation (LDA), which is a Bayesian topic modeling approach commonly used in natural language processing, with collapsed Gibbs sampler to classify chromatin regions into regulatory topics and classify and cluster cells based on their contributions to regulatory topics. Clustering of cells is achieved through optimising topic-cell distribution and classifying regions into topics is through region-topic distribution.

2.7 Benchmarking scATAC-Seq analysis tools

Currently, the interest in exploring the potential of scATAC-Seq booming than ever before. Furthermore, technologies including the commercially available ones such as 10X-ATAC make scATAC-Seq technologies readily available for even small laboratories. Therefore, a comprehensive assessment of existing analysis tools of scATAC-seq data and the development of pipelines that can suit the wide range of sequencing techniques will be critical. The only available benchmarking effort to-date is by Chen *et al.* (2019) [7].

Chen *et al.* (2019) [7] benchmarked the performance of 10 computational methods for scATAC-seq data analysis on different methods for generation of feature matrix and evaluated on clustering. They concluded that *SnapATAC* [1] performs consistently well across data sets of different sizes while and CisTopic [16] and methods developed by Cusanovich *et al.* (2018) [2] performed well in comparatively smaller datasets.

Chen *et al.* (2019) [7] also compared the differences between keeping and removing the first principal component (PC) when performing PCA dimension reduction. They presume that the first PC may only capture variation in sequencing depth. Corroborating with this presumption, methods that do not address sequencing depth, removing first PC will generally improve the clustering results. On the other hand, methods that implement binarisation (e.g. Cusanovich *et al.*(2018), *SnapATAC*) or that implement cell coverage bias correction (e.g. *chromVAR*, *SnapATAC*), tend to be less affected by sequencing depth, hence the removal of the first PC has no larger effect on clustering. Secondly, for methods using peaks as input, there was no clear difference between using bulk ATAC-seq peaks and peaks obtained from aggregated single cells. Only for *cisTopic*, Cusanovich *et al.*(2018), and Cicero did aggregating cell peaks perform better. Thirdly, after peak calling, *cisTopic*, *Scasat*, *SCRAT*, and *SnapATAC* filters out the ENCODE blacklist regions. The authors conclude that including this step does not have a major benefit on performance. Finally, for rare cell types, calling peaks using the pseudo-bulk of all cells may miss peaks specific for these rare cells. Without correction by pre-labelling these rare cells may not be identified [7].

2.7.1 Benchmark evaluation metrics

Chen *et al.* (2019) [7] used three commonly used metrices to evaluate the clustering results: a) the Adjusted Rand Index (ARI), b) Adjusted Mutual Information (AMI) and cluster homogeneity when ground truth (i.e. known cell group labels for the data, e.g. FACS labels) was available c) Residual Average Gini Index (RAGI) when ground truth was not available.

The ARI score is an adjusted version of Rand Index (RI) which measures similarity between two clusters. Mutual Information (MI) measures the mutual dependence between two clusters. Homogeneity measures whether the clustering algorithm assigns cells of the same class to each cluster. The Gini Index (GI) of a marker gene measures the imbalance in gene accessibility across clusters. GI is between 0 and 1, with 1 meaning imbalance i.e. a gene is accessible in only one cluster and 0 meaning a gene is accessible to an equal extent in both clusters. A positive RAGI value indicates that the marker gene separates the clusters better than a house keeping gene.

2.7.2 Benchmarking platform

Recently, our team developed *CellBench* [56], an R package containing functions and data structures that simplify the benchmarking of combinations of analysis methods without duplicating code. When using *CellBench*, methods are modified by wrapper functions to allow different methods taking a common input format and producing a common output format. It simplifies the process by generating many combinations from lists of methods in different analysis stages and allows individual combinations to fail without affecting the execution of the other method combinations. It also contains a function to time methods which can be useful for comparisons of running time.

3 Research Objectives

Single cell ATAC-seq has become a powerful technology for understanding cell population heterogeneity at the epigenetic level, with increasing throughput and easier data generation made possible through the commercial availability of kits. Interest in exploring scATAC-seq data is growing recently and to date a relatively small number of tailored computational analysis methods have been developed. Therefore, it is essential to benchmark these analysis methods in order to identify the key strengths and weaknesses of different solutions in search of a best method for each analysis step. Currently only one such benchmarking effort exists [7]. With currently available bioinformatics algorithms we can benchmark the existing tools with much intricacy and completeness. Benchmarking of existing tools in the current project will mainly focus on three broader aspects: 1) customisation of the workflow, 2) adjusting tool parameters and 3) using data with varying quality.

What will be covered:

1. This project aims to develop a universal performance indicators to evaluate the performance of current scATAC-seq analysis tools

2. This project also aims to evaluate the performance of current scATAC-seq analysis tools using developed evaluation metrices:

(a) on sequencing depth variation;

(b) on data quality variation;

(c) using default and customised parameters;

(d) in terms of the ease of use of the tool

using both experimental and simulated data sets which have known cell group labels available as ground truth.

What will be covered if time allows

1. Extension of the *scPipe* [57] package, which currently only handles scRNA-seq data, to also accommodate the preprocessing of raw scATAC-seq data may occur if time permits.

2. If access to working protocols can be obtained, this project may also generate new benchmarking data sets for single cell nucleosome, methylation and transcription sequencing (scNMT-seq).

3. A comparison of tools that are capable of combining single cell epigenomics (scATAC-seq) and transcriptomics (scRNA-deq) data from the same sample, which may further improve the clustering of cells, will also be considered if time allows.

4 Methodology

4.1 Available datasets

We have several in-house data sets as well as access to multiple public data sets.

4.1.1 In-house scATAC-seq and scRNA-seq data sets

An in-house scATAC-seq data set was generated on the 10x Chromium platform that is suitable for benchmarking. It contains an equal mixture of cells from five different human lung adenocarcinoma cell lines (HCC827, H1975, A549, H838 and H2228) (Figure 6) and the cell number is around 1,500. In addition to the scATAC-seq data set, single cell RNA-seq data sets for the same cell mixture were also generated, using two distinct protocols: CEL-seq2 and the 10x Chromium platform. These scRNA-seq data sets have been used for benchmarking scRNA-seq analysis methods as described in Tian *et al.* [56]. The availability of both scATAC-seq and scRNA-seq data can potentially be used for comparing methods that are capable of integrating both data types from the same sample.

In a preliminary analysis of the cell mixture scATAC-seq data using *Cell Ranger ATAC*, 7 clusters are identified within 1,320 cells. This may suggest epigenetic heterogeneity within some cell lines or that more detailed analysis may be needed.

4.1.2 Publicly available scATAC-seq data sets

Previously mentioned publicly available data sets for scATAC-seq will also be used. For example, the continuity of the Human hematopoietic lineage data set can be used for benchmarking trajectory analysis methods and the Mouse Cell Atlas data set can be used to benchmark the efficiency and scalability of different analysis methods.

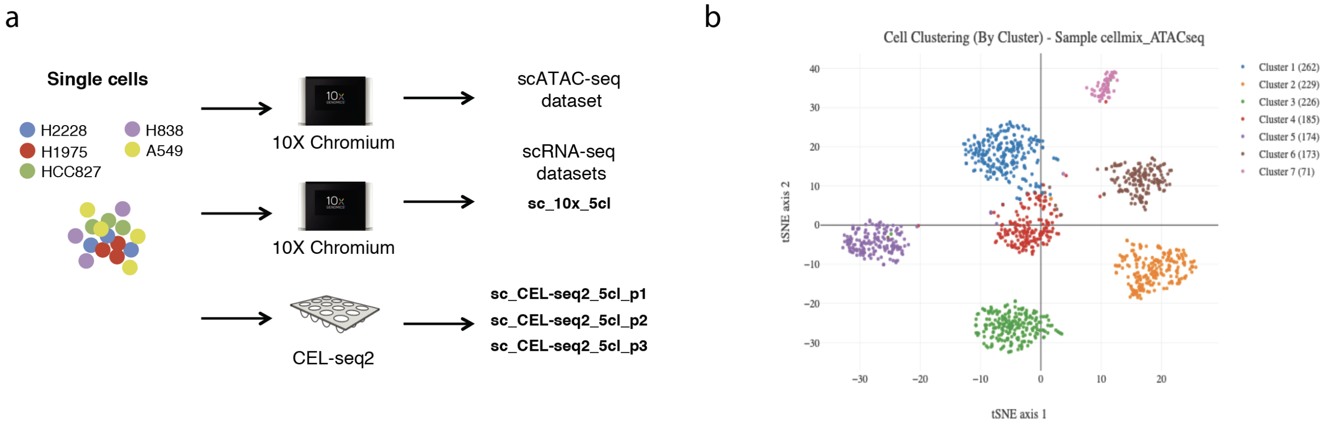

Figure 6: Summary of the in-house scRNA-seq and scATAC-seq data sets available for the planned benchmark analysis. Figure adapted from Tian *et al.* [56]. (a) Five human lung adenocarcinoma cell lines are mixed in equal proportions ($\sim$ 1,500 cells). 10X scATAC-seq and various different scRNA-seq protocols were performed on these cells. (b) Preliminary analysis of the cell mixture scATAC-seq data using *Cell Ranger ATAC*. A t-SNE plot of the Cell Ranger ATAC output identifies 7 clusters within 1,320 cells.

4.1.3 Simulated data: generating single cell ATAC-seq data from bulk ATAC-seq data

To simulate single-cell ATAC-seq data from bulk data sets, previous studies [7] have used the peak × cell type count matrix from bulk ATAC-seq and a binomial model. In detail, for a simulated single cell j having cell type k and peak i, the peak counts for this single cell is generated by $c_{i,j} \sim binom(2, p_i^t)$ where $p_i^t = (r_i^t)(\frac{1}{2}n)(1-q) + (\frac{1}{k})(\frac{1}{2}n)q$, with $q \in [0, 1]$ defining the noise level and n defining the number of simulated fragments. Some potential sources of publicly available bulk ATAC-seq data sets include: (1) human bone marrow bulk ATAC-seq (GSE119453), which contains FACS-sorted data from 6 cell types (HSCs, CMPs, erythroid cells (Ery), and three lymphoid cell types: NK, CD4 and CD8 T-cells) and (2) human erythropoiesis bulk ATAC-seq (GSE115672), which includes HSCs, CMPs, MEPs, MPPs, myeloid progenitors (MyP), colony forming unit-erythroid (CFU-E), proerythroblasts (ProE1), proerythroblasts (ProE2), basophilic erythroblasts (BasoE), polychromatic erytrhoblasts (PolyE), orthochromatic erythroblasts (OrthoE) and OrthoE and reticulocytes (Orth/Ret).

4.2 Development of metrics for benchmark evaluation

To evaluate clustering results, similar to Chen *et al.* (2019) [7], three commonly used metrics will be used: a) the Adjusted Rand Index (ARI), b) Adjusted Mutual Information (AMI) and cluster homogeneity when ground truth (i.e. known cell group labels for the data, e.g. FACS labels) are available and c) Residual Average Gini Index (RAGI) when ground truth is not available.

Evaluation metrics for other steps such as trajectory analysis and differential accessibility analysis are not currently well defined for scATAC-seq work. In this project, evaluation metrics for such tasks will be informed by those used in previous scRNA-seq analysis benchmarking efforts [56, 58].

4.3 Benchmarking scATAC-Seq tools

I plan to use the *CellBench* Bioconductor package in my benchmarking analysis to help streamline the processing and organise the results. Since *CellBench* is R-based, this project will mainly focus on benchmarking those analysis tools listed in Table 2 that are available as R packages.

4.3.1 Generating data sets with different noise levels and/or sequencing depth

To evaluate the performance of scATAC-seq tools where sequencing depth or different data quality varies, data sets with these features have to be generated. As described in section 4.1.3, Chen *et al.* (2019) [7] simulate data sets with different noise levels and sequencing depths from bulk ATAC-seq data. The starting point is an initial peak-by-cell-type matrix and the results are several new peak-by-cell matrices. This approach may not be ideal for benchmarking tools that do not use peaks as features. Therefore, coming up with simulation methods that start from the read level will be explored by sampling from reads from `fastq` files will be explored to benchmark feature matrix construction methods.

4.3.2 Evaluating the performance of scATAC-Seq tools using *CellBench*

While Chen *et al.* (2019) [7] only performed benchmarking on the feature matrix construction step, in this project I will also benchmark downstream analysis steps including clustering and trajectory analysis. The in-house data set will be used for benchmarking feature matrix generation and clustering but is less suitable for benchmarking trajectory analysis methods. The data sets generated from continuous cell types such as Human hematopoeitic cells can be used for benchmarking trajectory inference methods. *CellBench* would aid us to evaluate the performance of a host of R-based scATAC-Seq analysis tools by;

- exploring different feature matrix generation methods and downstream analysis methods with varied starting parameters
- enabling different combinations of matrix generation methods and downstream analysis methods

Finally, to evaluate the results, I will use the evaluation metrics developed during this project as mentioned in Section 4.2.

5 Research Plan

5.1 Semester 1, 2020

1. Become familiarised with the basics of Linux and R and begin exploring the analysis of scATAC-seq data using 2-3 methods mentioned, preferentially using comprehensive tools such as SnapATAC (∼ 1 month).
2. Evaluate the usability of each tool (∼ 2 weeks).
3. Benchmark the analysis tools with the help of *CellBench* (∼ 2 months).
4. In the last 1.5 months of the semester, start to write the progress report.

5.2 Semester 2, 2020

1. Benchmark a further 3-4 analysis tools with the help of *CellBench* (∼ 1 months).

2. Draw conclusions from the benchmarking work (∼ 2 weeks)

3. In the last 2 months of the semester, start to write the progress report and prepare for the final presentation.

References

1. Fang, R. *et al.* Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types. *bioRxiv,* 1–41 (2019).

2. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174,** 1309–1324. ISSN: 10974172 (2018).

3. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523,** 486–490. ISSN: 14764687 (2015).

4. Mezger, A. *et al.* High-throughput chromatin accessibility profiling at single-cell resolution. *Nature Communications* **9,** 6–11. ISSN: 20411723 (Dec. 2018).

5. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology* **37,** 925–936. ISSN: 1087-0156 (2019).

6. Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nature Biotechnology* **37.** ISSN: 1087-0156. doi:10.1038/s41587-019-0147-6. http://www.nature.com/articles/s41587-019-0147-6 (2019).

7. Chen, H. *et al.* Assessment of computational methods for the analysis of single-cell ATAC-seq data. *bioRxiv,* 739011 (Aug. 2019).

8. Yu, W., Uzun, Y., Zhu, Q., Chen, C. & Tan, K. scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data. *bioRxiv,* 824326 (Oct. 2019).

9. Ji, Z., Zhou, W. & Ji, H. Single-cell regulome data analysis by SCRAT. *Bioinformatics* **33,** 2930–2932. ISSN: 14602059 (2017).

10. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods* **14,** 975–978. ISSN: 15487105 (2017).

11. Zamanighomi, M. *et al.* Unsupervised clustering and epigenetic classification of single cells. *Nature Communications* **9,** 1–8. ISSN: 20411723 (2018).

12. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Molecular Cell* **71,** 858–871. ISSN: 10974164 (2018).

13. De Boer, C. G. & Regev, A. BROCKMAN: Deciphering variance in epigenomic regulators by k-mer factorization. *BMC Bioinformatics* **19,** 1–13. ISSN: 14712105 (2018).

14. Baker, S. M., Rogerson, C., Hayes, A., Sharrocks, A. D. & Rattray, M. Classifying cells with Scasat, a single-cell ATAC-seq analysis tool. *Nucleic acids research* **47,** e10. ISSN: 13624962 (2019).

15. Urrutia, E., Chen, L., Zhou, H. & Jiang, Y. Destin: toolkit for single-cell analysis of chromatin accessibility. *Bioinformatics.* ISSN: 1367-4803. doi:10.1093/bioinformatics/btz141 (Mar. 2019).

16. Bravo González-Blas, C. *et al.* cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature Methods* **16,** 397–400. ISSN: 15487105 (May 2019).

17. Lal, A. *et al.* AtacWorks: A deep convolutional neural network toolkit for epigenomics. *bioRxiv,* 829481 (Nov. 2019).

18. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389,** 251–260. ISSN: 00280836 (1997).

19. Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423,** 145–150. ISSN: 00280836 (May 2003).

20. Kornberg, R. D. Chromatin structure: A repeating unit of histones and DNA. *Science* **184,** 868–871. ISSN: 00368075 (1974).

21. Henikoff, S. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nature Reviews Genetics* **9,** 15–26. ISSN: 14710056 (Jan. 2008).

22. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nature Reviews Genetics* **17,** 487–500. ISSN: 14710064 (Aug. 2016).

23. Dann, G. P. *et al.* ISWI chromatin remodellers sense nucleosome modifications to determine substrate preference. *Nature* **548,** 607–611. ISSN: 14764687 (Aug. 2017).

24. Lee, C. K., Shibata, Y., Rao, B., Strahl, B. D. & Lieb, J. D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics* **36,** 900–905. ISSN: 10614036 (Aug. 2004).

25. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82. ISSN: 00280836 (Sept. 2012).

26. Radman-Livaja, M. & Rando, O. J. Nucleosome positioning: How is it established, and why does it matter? *Developmental Biology* **339,** 258–266. ISSN: 1095564X (Mar. 2010).

27. Tsompana, M. & Buck, M. J. Chromatin accessibility: A window into the genome. *Epigenetics and Chromatin* **7.** ISSN: 17568935. doi:10.1186/1756-8935-7-33 (2014).

28. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* **20,** 207–220. ISSN: 1471-0064 (Apr. 2019).

29. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132,** 311–22. ISSN: 1097-4172 (Jan. 2008).

30. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods* **6,** 283–9. ISSN: 1548-7105 (Apr. 2009).

31. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10,** 1213–1218. ISSN: 15487091 (2013).

32. Mieczkowski, J. *et al.* MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nature communications* **7,** 11485. ISSN: 2041-1723 (2016).

33. Mueller, B. *et al.* Widespread changes in nucleosome accessibility without changes in nucleosome occupancy during a rapid transcriptional induction. *Genes & development* **31,** 451–462. ISSN: 1549-5477 (2017).

34. Kelly, T. K. *et al.* Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Research* **22** (Dec. 2012).

35. Cusanovich, D. A. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348,** 910–914. ISSN: 10959203 (May 2015).

36. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics.* ISSN: 15461718. doi:`10.1038/ng.3646` (2016).

37. Cusanovich, D. A. *et al.* The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555,** 538–542. ISSN: 14764687 (Mar. 2018).

38. Preissl, S. *et al.* Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nature Neuroscience* **21,** 432–439. ISSN: 15461726 (Mar. 2018).

39. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173,** 1535–1548. ISSN: 10974172 (2018).

40. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33,** 495–502. ISSN: 15461696 (May 2015).

41. Andrews, S. *FastQC: A quality control tool for high throughput sequence data.* 2010.

42. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10 (May 2011).

43. Krueger, F. Trim Galore! `https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/` (2012).

44. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760. ISSN: 1367-4803 (July 2009).

45. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359. ISSN: 15487091 (Apr. 2012).

46. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079. ISSN: 13674803 (Aug. 2009).

47. 10X Genomics. Cell Ranger ATAC (2018).

48. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9,** R137. ISSN: 1465-6906 (2008).

49. Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W. & Lieb, J. D. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biology* **12.** ISSN: 14747596. doi:10.1186/gb-2011-12-7-r67 (July 2011).

50. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R Journal* **8,** 289–317. ISSN: 20734859 (2016).

51. Studer, M. WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R. *LIVES Working Papers* **24.** doi:http://dx.doi.org/10.12682/lives.2296-1658.2013.24 (2013).

52. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008.** ISSN: 17425468. doi:10.1088/1742-5468/2008/10/P10008 (Oct. 2008).

53. Van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605 (2008).

54. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. http://arxiv.org/abs/1802.03426 (Feb. 2018).

55. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. ISSN: 14764687 (Sept. 2012).

56. Tian, L. *et al.* Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature Methods* **16,** 479–487. ISSN: 15487105 (2019).

57. Tian, L. *et al.* scPipe: A flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. *PLoS Computational Biology.* ISSN: 15537358. doi:10.1371/journal.pcbi.1006361 (2018).

58. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nature Biotechnology* **37,** 547–554. ISSN: 15461696 (2019).