

Towards Understanding Genome Regulation via High-resolution Analysis of Chromatin Accessibility

Liangti Dai

Linacre College
University of Oxford

*A thesis submitted for the degree of
Doctor of Philosophy*

Trinity 2023

Abstract

Next generation sequencing has been used to in functional genomics to represent specific aspects of chromatin structure, DNA-protein interactions and the epigenome, enriching the knowledge of the non-coding genome which plays significantly roles in gene regulation but is vaguely understood. Among the assays for functional genomics, Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) is a powerful and popular tool as its simple read out of ‘chromatin accessibility’ provides both the positioning of functional regulatory elements and a general representation of the active genome. Despite its advantages, the analysis of ATAC-seq is challenging and due to the data sparsity and the sub-optimal use of the data, especially the fragment size of the sequencing reads. In this thesis, I address the use of ATAC-seq fragment size in two different ways with respect to different goals: prioritising functional variants and peak calling.

The first goal is achieved through a collaborative work in AVOCATO, which develops an end-to-end pipeline for single-cell ATAC-seq (scATAC-seq) functional analysis. AVOCATO employs a simple but useful fragment size filtering strategy, retaining only short fragments from high-quality scATAC-seq data, which successfully prioritises functional SNP variants hidden in the raw signals. Several statistical optimisations in pre-processing stages and a powerful interactive visualisation platform assist AVOCATO to be a solid and hands-on tool for high-resolution scATAC-seq analysis.

The second goal, peak calling, focuses on both bulk and single-cell ATAC-seq data. I developed EpiCall, a novel ATAC-seq specific peak caller that shows superior performance compared to current popular peak callers. By modeling ATAC-seq fragments of varying sizes differently, EpiCall optimally uses fragment size and coverage information, which is integrated into a mechanistic model for the actual formation of sequencing fragments. I applied EpiCall in various datasets and evaluated it extensively with current peak callers. For bulk ATAC-seq datasets, I demonstrate the precision and recall between peak calls and ENCODE functional annotations, as well as the enrichment of motifs and histone modifications. For scATAC-seq datasets, I show the ability of EpiCall to improve differential motif enrichment across single cells and cell clusters, and in detecting and distinguishing neighborhood peaks sensitively. For both bulk and single-cell datasets, I show the robustness of EpiCall in peak calling down-sampled low-quality data, a significant advantage over current peak callers.

Towards Understanding Genome Regulation via High-resolution Analysis of Chromatin Accessibility



Liangti Dai
Linacre College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity 2023

Acknowledgements

I am fortunate to have received great help from many people during my PhD journey. First and foremost, I am deeply grateful to my supervisors, Professor Gerton Lunter and Professor Jim Hughes, for their knowledge, support and constant encouragement. As an expert in mathematics and computational genetics, Gerton has been providing solid and efficient guidance throughout my PhD. I wouldn't have completed this thesis without his input on the theoretical foundation of EpiCall. Moreover, his optimistic and peaceful mind cheered me up during the change of my research projects and through all the ups and downs. Jim has offered me enormous help especially on the application side of method development since the start of the pandemic. As an encyclopedia of functional genomics and immunology, he inspired most of the evaluation criteria for EpiCall and guided me towards the big picture while I was drowning into the details. I am more than grateful for his invaluable suggestions during my thesis writing.

I would like to express my sincere gratitude to everyone in the Lunter group and Hughes group. Many thanks to Richard Brown, Daniel Cooke, Ron Schwessinger, Christopher Cole, Emine Ravza Gür, and Edward Sanders for making the Lunter group inclusive and welcoming when I first joined (and became unexpectedly the last student of the group). I also want to acknowledge everyone in the Hughes group for your inspiring, constant passion for research and life. I am honored to work on the Avocado project in collaboration with Chris, Ravza, Edward, Martin Sergeant and Simone Riva along with the enjoyable casual chats during this time.

The MPLS Doctoral Training Centre (DTC) and my PhD program Interdisciplinary Bioscience DTP have been a warm home base ever since the start of this journey. In addition to the complete freedom to explore for my research direction, I have been benefited from the DTC networking to build many valuable connections with current students, alumni, and faculty. Specifically, I would like to thank Dr. Ben Lambert, who initiated and organized the CoMo exploration and public engagement across DTC, for coordinating a collaboration with fellow students outside of my research domain that offered me great fun and a tiny taste of epidemiology.

I would also like to thank the Radcliffe Department of Medicine and MRC Weatherall Institute of Molecular Medicine (WIMM) for the wonderful research space and institutional support. I would like to acknowledge the Centre for Computational

Biology at WIMM and Wellcome Centre for Human Genetics for the essential support of high performance computing facilities CBRG and BMRC.

Words cannot express my deepest emotions to my dear family and old friends. Although the pandemic separated us apart physically for three years, knowing that you are always there provides me with the strongest foundation. Finally, a special thank to my partner, and best wishes for your own thesis writing.

Contents

List of Figures	vii
List of Tables	ix
List of Abbreviations	x
1 Functional Genomics and Next-generation Sequencing	1
1.1 The Non-coding Genome: Complex and Crucial	3
1.2 Next-Generation Sequencing in Profiling Functional Genomics	5
1.2.1 History of Nucleotide Sequencing Technologies	6
1.2.2 NGS Assays for Functional Genomics	7
1.3 ATAC-seq: Simple Assay with Rich Functional Information	8
1.3.1 Experimental Workflow of ATAC-seq	9
1.3.2 Standard Analysis of ATAC-seq	9
1.4 Single-cell ATAC-seq (scATAC-seq)	12
1.4.1 scATAC-seq Experimental Approaches	12
1.4.2 Standard Analysis of scATAC-seq	13
1.5 ATAC-seq Peak Calling	16
1.5.1 Challenges of ATAC-seq Peak Calling	18
1.5.2 Current ATAC-seq Peak Calling Algorithms	18
1.6 Multiple levels of Gene Regulation Represented in ATAC-seq Fragment Size	20
1.7 Machine Learning Applications in Functional Genomics	21
1.7.1 Identifying and Annotating Functional Elements	22
1.7.2 Signal Dnoise, Dropout Imputation and Batch Corrections	22
1.7.3 Functional Predictions	23
1.8 Aims and Scope of the Thesis	23
2 AVOCATO: A High-resolution Analysis Pipeline for scATAC-seq	
Data	26
2.1 Introduction	26
2.2 Results	27
2.2.1 The Characteristics of AVOCATO	27

2.2.2	The Functionalities of the Workflow	28
2.2.3	Configuration, Execution and Visualisation	30
2.2.4	Example: Visualisation of 10X PBMC 3K scATAC-seq data analysis	31
2.3	Methods	33
2.3.1	Whitelist-free barcode correction	33
2.3.2	Adapting ArchR for Stage 1 Analysis	35
2.3.3	Iterative Decision Tree for Selecting Cluster-Representative Genes	36
2.4	Code Availability	36
2.5	Acknowledgement	36
3	EpiCall: A High-Resolution Peak Caller for ATAC-seq Data	39
3.1	Introduction	40
3.2	EpiCall	43
3.2.1	The Model Concept	43
3.2.2	Derivation of the Probabilistic Model	43
3.2.3	Implementation With hidden Markov Model	47
3.2.4	Calling Peaks With the Viterbi Algorithm	48
3.2.5	Defining Peak Score	48
3.3	Results	48
3.3.1	Accurate and sensitive detection of cell type-specific open chromatin regions	49
3.3.2	EpiCall sensitively recovers TF binding and functional histone modification signals	53
3.3.3	High Evolutionary Conservation in EpiCall Peaks	59
3.3.4	Robustness to low sequencing depth	61
3.3.5	Integrating fragment size and coverage benefits peak calling	62
3.4	Discussion	66
3.4.1	High-resolution peak calls enable accurate signal prioritisation of epigenomic profiling	66
3.4.2	Future work	66
3.5	Code and Data Availability	69
3.6	Methods	69
3.6.1	Data Preprocessing	69
3.6.2	Peak Calling	69
3.6.3	AUPRC measurement of ATAC-seq peak calls	71
3.6.4	Motif enrichment analysis	72
3.7	Acknowledgement	72

4	EpiCall Iterative Peak Caller Facilitates Differential Regulatory Analysis Using scATAC-seq	73
4.1	Introduction	74
4.2	Results	75
4.2.1	Robustness of Peak Calling on Low Cell-number scATAC-seq Datasets	75
4.2.2	Differential Motif Enrichment Analysis for <i>Buenrostro2018</i> Dataset	76
4.2.3	Improved Cell Representation and Differential Regulatory Motifs analysis for 10X PBMC10K Multiomics Dataset	80
4.2.4	EpiCall Sensitive Recovers TF and Enhancer Signals from Proerythroblasts scATAC-seq Data	82
4.3	Discussion	86
4.3.1	Peak Calling for Rare Cells	86
4.4	Methods	87
4.4.1	scATAC-seq Data Pre-processing	87
4.4.2	Differential Motif Enrichment Analysis	88
4.4.3	PBMC10K Cell Type Annotations	88
4.4.4	Evaluation for Cell Clustering with Ground Truth	89
4.5	Acknowledgment	89
5	Conclusion	91
5.1	Summary	92
5.1.1	High-resolution scATAC-seq analysis aids the prioritisation of functional variants	92
5.1.2	A mechanistic model as a robust and sensitive ATAC-seq peak caller	92
5.2	Better functional annotations in need for ATAC-seq peak calling evaluation	93
5.3	Single cell representation learning: dimension reduction and cell state characterisation	95
5.4	Towards systematic benchmarking and setting up scATAC-seq analysis paradigms	97

Appendices

References	106
-------------------	------------

List of Figures

1.1	NGS assays and the functional genomics they profile.	5
1.2	Signal coverage of ATAC-seq in comparison with other epigenome assays.	9
1.3	Overview of ATAC-seq experimental protocols.	10
1.4	The standard analysis pipeline for bulk ATAC-seq data.	11
1.5	scATAC-seq Analysis Workflow.	14
1.6	The fragment size of ATAC-seq data holds rich information of functional genomics.	17
2.1	The End-to-end Workflow of AVOCATO.	30
2.2	The graphical user interface of AVOCATO.	32
2.3	The visualisation platform of AVOCATO demonstrating 10X PBMC3k scATAC-seq data analysis.	34
3.1	EpiCall model structure.	42
3.2	Hyper-parameter optimisation for EpiCall using GSE47753 GM12878 50K as training data.	50
3.3	Precision-Recall curves (PRC) of peak calls using four peak callers across five bulk ATAC-seq datasets.	52
3.4	Pileup average signal of evolutionary conservation and epigenomics surrounding peak centers or summits.	57
3.5	EpiCall produces bulk ATAC-seq peak calls sensitively correlated to functional epigenomics.	58
3.6	Epigenomics signals surrounding top-scored exclusive peak calls produced by EpiCall (A), MACS2 (B) and LanceOtron (C) on GM12878 in chromosome 1.	60
3.7	Peak calling performance on read down-sampled GM12878 bulk ATAC-seq evaluated by ‘self comparison’ (left) and ‘general performance’ (right).	63
3.8	K562 ATAC-seq peak calls in reference with functional genomics assays surrounding the hemoglobin gene locus HBB and HBG subunits, showing potential true and false positive peak calls.	65

4.1	Peak calling performance on down-sampled 10X GM12878 scATAC-seq evaluated by ‘self comparison’ (left) and ‘general performance’ (right).	77
4.2	Schematic view of human hematopoiesis with emphasis on erythropoiesis.	78
4.3	Summary of the <i>Buenrostro2018</i> dataset.	79
4.4	chromVAR differential motif enrichment of representative motifs on <i>Buenrostro2018</i>	81
4.5	Cell type annotations, cell clustering performance and chromVAR differential motif enrichment of representative motifs on <i>10X PBMC 10K multiome</i> dataset.	83
4.6	scATAC-seq proerythroblast at KLF1 locus, showing correspondence with ChIP-seq coverage of TF and Histone modifications.	85
4.7	scATAC-seq proerythroblast around SLC25A37 locus, showing correspondence to ChIP-seq coverage of TF and Histone modifications.	86
5.1	ChromHMM annotations are sub-optimal for benchmarking ATAC-seq peak calls.	95
2	Signals of functional genomics surrounding the center of EpiCall and LanceOtron peak calls on GM12878 50k dataset.	101
3	Signals of functional genomics surrounding the center and summit of MACS2 peak calls on GM12878 50k dataset.	102
4	Signals of functional genomics surrounding the center and summit of HMMRATAC peak calls on GM12878 50k dataset.	103

List of Tables

3.1	Summary statistics of peak calls across five cell lines.	53
3.2	Motif enrichment statistics of ATAC-seq peak calls in cell lines GM12878, K562 and HepG2.	55
1	Source of data used in this thesis	104
2	18-state ChromHMM Annotations	105

List of Abbreviations

NGS		next-generation sequencing
ATAC-seq	. . .	The Assay for Transposase-Accessible Chromatin with Sequencing
CREs		<i>cis</i> -regulatory elements
TFs		transcription factors
TSS		transcription starting sites
GWAS		genome-wide association study
NFRs		nucleosome-free regions
PCR		polymerase chain reaction
dNTPs		deoxynucleotide triphosphates
ddNTPs		di-deoxynucleotide triphosphates
SNP		single-nucleotide polymorphism
ChIP-seq	. . .	Chromatin Immunoprecipitation Sequencing
MNase-seq	. . .	Micrococcal Nuclease sequencing
FAIRE-seq	. . .	Formaldehyde-Assisted Isolation of Regulatory Elements sequencing
DNase-seq	. . .	DNase I hypersensitive sites sequencing
3C		Chromosome Conformation Capture
MAPQ		Mapping Quality
QC		Quality Control
scRNA-seq	. . .	single-cell RNA-seq
scATAC-seq	. . .	single-cell ATAC-seq
scHi-C		single-cell Hi-C
CITE-seq	. . .	cellular indexing of transcriptomes and epitopes
FACS		fluorescence-activated sorting
IFC		integrated fluidics circuit

GEM	gel bead-in-Emulsion
LSI	latent semantic indexing
LDA	Latent Dirichlet Allocation
NMF	non-negative matrix factorization
VAE	variational auto-encoders
CNN	convolutional neural networks
cCREs	candidate <i>cis</i> -regulatory elements
HMM	hidden Markov Model
Avocato	Analysis and Visualization Of Single-Cell ATAC-seq Observations
UMAP	Uniform Manifold Approximation and Projection
iLSI	iterative Latent Semantic Indexing
<i>k</i>-NN	<i>k</i> -nearest neighbors
ROC	receiver operating characteristic curve
AUROC	area under the ROC
AUORC	area under the precision-recall curve
PWMs	position weight matrices
TFBS	transcription factor binding sites
ARI	adjusted Rand index
AMI	adjusted Mutual Information
SVD	Singular Vector Decomposition
SVM	Support Vector Machines
NLP	Natural Language Processing

1

Functional Genomics and Next-generation Sequencing

Contents

1.1	The Non-coding Genome: Complex and Crucial	3
1.2	Next-Generation Sequencing in Profiling Functional Genomics	5
1.2.1	History of Nucleotide Sequencing Technologies	6
1.2.2	NGS Assays for Functional Genomics	7
1.3	ATAC-seq: Simple Assay with Rich Functional Information	8
1.3.1	Experimental Workflow of ATAC-seq	9
1.3.2	Standard Analysis of ATAC-seq	9
1.4	Single-cell ATAC-seq (scATAC-seq)	12
1.4.1	scATAC-seq Experimental Approaches	12
1.4.2	Standard Analysis of scATAC-seq	13
1.5	ATAC-seq Peak Calling	16
1.5.1	Challenges of ATAC-seq Peak Calling	18
1.5.2	Current ATAC-seq Peak Calling Algorithms	18
1.6	Multiple levels of Gene Regulation Represented in ATAC-seq Fragment Size	20
1.7	Machine Learning Applications in Functional Genomics	21
1.7.1	Identifying and Annotating Functional Elements	22
1.7.2	Signal Dnoise, Dropout Imputation and Batch Corrections	22
1.7.3	Functional Predictions	23
1.8	Aims and Scope of the Thesis	23

The long-term goal of genomics is to decipher the mechanisms governing the

functions of the human body both on the molecular and the systemic level: to reveal the sophisticated regulatory programs responsible for which genes are expressed when, where, and how in the various tissues, or in short, *gene regulation*. Gene regulation comprises mechanisms that are present in all life forms, from prokaryotes to eukaryotic species, and drive almost all cellular activities from the most basic responses to extracellular stimuli to the development from a single cell into various tissue types in multi-cellular systems.

Since being proposed in the 1950s, the central dogma, which states that genes code for proteins via messenger RNA, has guided genomics and molecular biology to focus on the *coding* genome. The fully decoded mappings between nucleotides, codons and amino acids directly link genetic variants with the resulting amino acid changes, which subsequently bring about phenotypic changes. However, the fact that only 1.5% of the genome are protein-coding regions brings the question of what functions lie in the remaining 98.5% *non-coding* genome. Although much of the human genome is so-called ‘junk’ DNA, comparative genomic analyses indicate that as much as 8% is under selective constraint and therefore functional [1]. Following the completion of the Human Genome Project in 2003 [2], technical developments have been expanding the scale, quality and variety of genome-wide sequencing covering both the coding and non-coding genome at lower cost every year. This unprecedented access to the DNA sequence underpinning all of our biology brought with it a plethora of opportunities for functional, evolutionary and computational analyses. Many of these required the development of novel and bespoke computational methods to accurately process and model these data in order to analyse, discover and explain biology from these ‘big data’ and moving towards the reconstruction of the gene regulatory networks, making connections between non-coding and the coding genome.

In this chapter, I will introduce the importance of the non-coding genome in understanding functional genomics, the applications of next-generation sequencing (NGS) in various genomics profiling and the open challenges in the field in terms of technological limitations and statistical analysis. In particular, I will emphasize

on ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing), the sequencing data this thesis is focused on. ATAC-seq serves a superior assay to study the non-coding genome as its simple read out of ‘chromatin accessibility’ reveals multiple aspects of gene regulation [3], reflecting both the *positioning* of functional regulatory elements and a *general* representation of the active genome, two important facets that most other assays are not able to reflect both. Despite all the advantages, the analysis of ATAC-seq remains challenging due to the sub-optimal use of the data, in particular the *fragment size* of the sequencing reads. I will briefly introduce the different usages of the ATAC-seq fragment size information with respect to two different goals: prioritising functional variants related to transcription factor binding, which is extended in Chapter 2, and developing a novel ATAC-seq-specific peak caller by integrating fragment size with sequencing coverage to create a mechanistic model on ATAC-seq fragment formation, which is thoroughly introduced in Chapter 3-4.

1.1 The Non-coding Genome: Complex and Crucial

The non-coding genome regulates gene expression by controlling chromatin structure, the recruitment of transcription factor complexes, gene transcription, all the way to post-transcriptional RNA splicing and modification, and finally translation. One category of functional non-coding elements that play an important role in this process are known as *cis*-regulatory elements (CREs), which refer to regions in the genome with clusters of specific DNA sequences (termed *motifs*) that are recognisable by DNA-binding proteins such as transcription factors (TFs). Depending on their relative locations to gene bodies and their role in gene regulation, they can be categorised into promoters, enhancers, silencers and insulators. *Promoters* sit prior to the transcription starting sites (TSS) at the 5’ end of gene bodies and contain highly conserved sequences, and act to recruit TFs and RNA polymerase II to eventually commence gene transcription. Other CREs are involved in transcription activation (*enhancers*) and repression (*silencers*), and physical blockage of regulatory influence (*insulators*) in more intricate ways and across a range of distances, up

to several megabase (Mb) along the linear genome from TSSs. On top of the linear relations to genes, the chromosome is folded into spatial structures guided by insulator (CTCF) elements, enhancer interactions, histone proteins and other complex mechanisms, and these structures promote or repress physical long-range interactions of regulatory elements. These include ‘super enhancers’, regions that span tens of kilobases and are densely packed with enhancer motifs and other regulatory elements, which act to regulate gene expression efficiently by contacting with target promoters through direct chromatin looping [4, 5].

Thanks to worldwide collaborative projects such as ENCODE [6, 7] and Roadmap Epigenomics [8–10] over the past several two decades, researchers have partly characterised the crucial role that the non-coding genome plays in gene regulation. At the same time, population genomics assays such as genome-wide association studies (GWAS) and population-wide projects such as 1000 Genomes [11] have identified many disease-associated genetic variants. The majority of these are located in the non-coding genome, drawing further attention to non-coding regulatory elements and the need to understand how they affect gene regulation [12, 13].

To regulate genes, the DNA motifs that form *cis*-regulatory elements need to be physically accessible to TFs; conversely, some TFs are hypothesized to act as pioneer factors that are themselves able to make chromatin accessible [14]. As a result, genome-wide chromatin accessibility assays are highly informative about cell-specific activity of regulatory elements. Figure 1.1 shows an overview of the connections between the eukaryotic chromatin structure, various CREs and the epigenomics sequencing assays to profile them. As shown, the linear DNA wraps around the histone proteins to form nucleosomes, which condense into chromatin and finally super-coil into chromosomes, forming different levels of condensation ranging from loosely packed euchromatin and highly compacted heterochromatin [15]. Euchromatin often contains sites that are hypersensitive to DNase I, an enzyme that cleaves DNA loci that are less occupied by histone proteins and nucleosomes (i.e. nucleosome-free regions, NFRs). This is a proxy for sites where the CREs have higher chance to be accessible to TFs to initiate and activate gene transcription.

By contrast, sites that are less sensitive to DNase-I are hidden in heterochromatin and are more likely to be inactive.

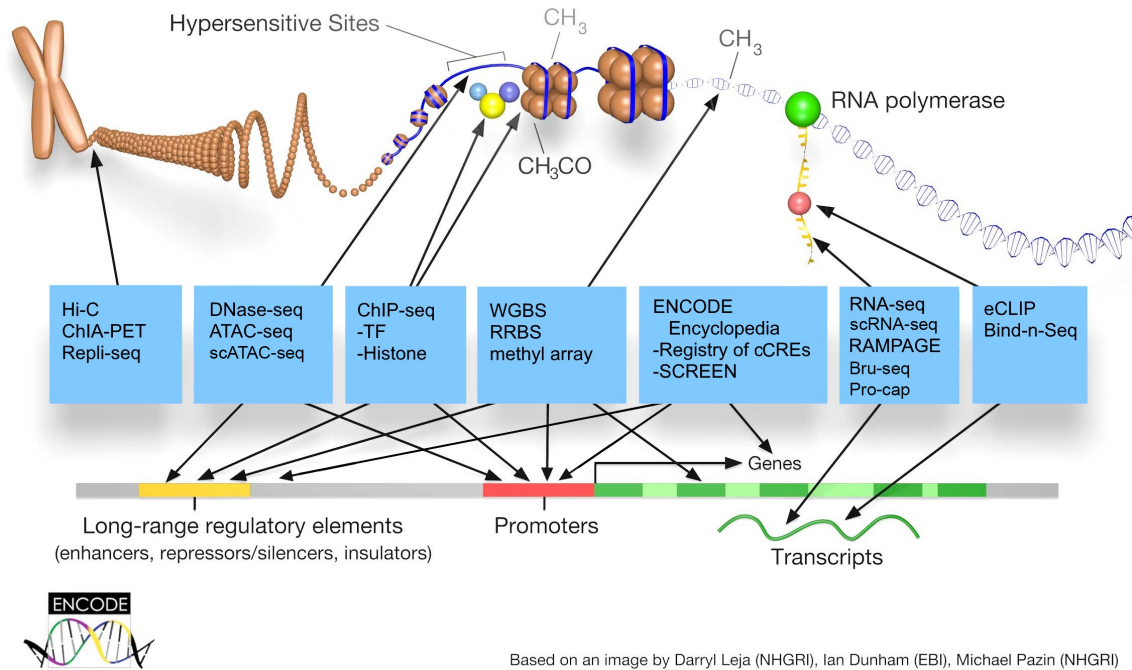


Figure 1.1: NGS assays and the functional genomics they profile. Showing correspondence to different stages of gene regulation. Figure from <https://screen.wenglab.org/> [16].

1.2 Next-Generation Sequencing in Profiling Functional Genomics

Following the discovery of the DNA double-helix structure in 1950s, efforts have been made to determine the order and identity of bases along DNA molecules and the whole genome of organisms (i.e. base calling or base detection), and progressively extending to more specific sequencing aims such as the profiling of exome, transcriptome, epigenome and DNA-protein interactions (section 1.2.2). The sequencing techniques involved can be categorised into three stages according to their logic and chemistry: first-generation sequencing, NGS and long-read sequencing.

I will begin this section with a brief introduction of the sequencing technology evolution with emphasis on NGS, followed by a few representative bulk and single-

cell NGS sequencing applications in functional genomics, as the majority of raw data related in this thesis belongs to NGS.

1.2.1 History of Nucleotide Sequencing Technologies

First-generation sequencing: One of the earliest sequencing approach is Sanger sequencing developed in the 1970s, famous for the use of chain-termination in base calling. In Sanger sequencing, DNA molecules are amplified using sequence-specific primers and the polymerase chain reaction (PCR), which appends single bases of deoxynucleotide triphosphates (dNTPs) sequentially to the growing DNA copy. To identify the bases, Sanger mixed normal dNTPs with modified di-deoxynucleotide triphosphates (ddNTPs), utilising the fact that ddNTPs terminate the elongation of DNA strand. In addition, the ddNTPs are labelled radioactively or fluorescently beforehand. By separating the incomplete fragments by size using gel electrophoresis, the DNA sequence can then be determined. Sanger sequencing was used to sequence the first whole genomes of several organisms, and was the major method used in the Human Genome Project. Today, it still stands out in terms of base calling accuracy and is widely applied in short-length sequencing and clinical projects. The basic workflow of Sanger sequencing, *fragmentation* $\rightarrow$ *PCR amplification* $\rightarrow$ *detection (base calling)*, remains the backbone of many subsequent approaches.

Next-generation sequencing: First-generation sequencing is known to be relatively costly and time consuming, and is suited mainly for short sequences. The sheer size of the human genome sequencing project drove the development of faster, higher-volume and more cost-effective technologies. Since early 2000s, NGS have been developed to realise low-cost high-throughput sequencing in which vast numbers of short reads are sequenced in a single experiment.

The workflow of NGS can be summarized as *fragmentation* $\rightarrow$ *adaptor ligation* $\rightarrow$ *library preparation* $\rightarrow$ *PCR amplification* $\rightarrow$ *base calling*. The Current NGS platform is dominated by Illumina's (Solexa) NextSeq/NovaSeq platform [17, 18], a 'Sequencing by Synthesis' approach that identifies DNA bases by their unique fluorescent signals and simultaneously adding them to the nucleic acid chain. The

PCR amplification is carried out using a bridge-amplification approach which allows massively parallel sequencing over different fragments clustered to each flow cell channel [17–19].

Benefiting from the high-throughput production of short reads (50-200bp), NGS assays are excellent for resequencing projects, SNP (single-nucleotide polymorphism) calling, and targeted sequencing of very short amplicons. The flexibility of NGS has also made it the mainstream sequencing platforms and the basis of many new variations in profiling functional genomics.

Third/fourth generation sequencing: The short reads produced by NGS, limited by their lengths, can lead to sequence gaps and incomplete assemblies, thus are not suitable for all sequencing purposes. To overcome the limitation, long-sequencing technologies are developed for highly accurate real-time single-molecule DNA sequencing. Successful methods include Single Molecule, Real-Time (SMRT) sequencing developed by PacBio and nanopore sequencing based on pore-forming proteins and electrical detection developed by Oxford Nanopore Technologies that supports sequencing lengths up to millions of bases [20]. These methods have been particularly fit for *de novo* genome sequencing, phasing and detection of structural variants without the need for assembly.

1.2.2 NGS Assays for Functional Genomics

Today, NGS assays are used far beyond its initial use, that of sequencing nuclear DNA. Numerous functional assays have been built around the technology, probing at different targets to represent specific aspects of chromatin structure and the epigenome. As shown in Figure 1.1, RNA-seq profiles *transcriptome*; Chromatin Immunoprecipitation Sequencing (ChIP-seq) targeting of TF binding profiles *DNA-protein interactions*; ChIP-seq targeting of histone modifications profiles *histone modifications associated with active or repressed gene transcription*; Micrococcal Nuclease sequencing (MNase-seq), Formaldehyde-Assisted Isolation of Regulatory Elements sequencing (FAIRE-seq), DNase-I hypersensitive sites sequencing (DNase-seq) and ATAC-seq, all profile the accessibility of chromatin. Hi-C, Capture-C

and numerous other Chromosome Conformation Capture (3C) method assay the dynamic interactions of the genome in 3D space.

However, most assays only cover a single specific aspect of the epigenome, requiring many experiments to provide a full picture. For example, a single ChIP-seq experiment focuses only on the binding regions of a certain TF or histone marks while missing the others; RNA-seq targets at the transcriptome that ignores the non-coding genome and the participation of regulatory elements. In contrast, assays for chromatin accessibility give genome-wide information on both the functional positioning and a general measure of regulation ‘activity’. They represent perhaps the single most informative assay reflecting the multi-level complex epigenomics landscape including accessibility, TF and histone binding, nucleosome positioning and characterisation of active CREs.

1.3 ATAC-seq: Simple Assay with Rich Functional Information

Among the epigenome assays targeting genome-wide chromatin accessibility, ATAC-seq has stood out since its emergence in 2013 because of the simple and fast experimental protocol, high sensitivity and robustness in signal detection, and the potential to read out multidimensional and simultaneous signals on chromatin accessibility, nucleosome positioning and TF footprinting [21]. Signals of ATAC-seq align in the correspondence to epigenomics assays such as ChIP-seq of CTCF (promoter-binding motifs), transcription factors, H3K4me1 (histone methylation primarily associated with enhancers), H3K4me3 (histone methylation primarily associated with promoters) and H3K27ac (histone acetylation associated with both enhancers and promoters) (Figure 1.2).

In this section, I will briefly introduce the experimental workflow, the standard data processing and the analytical challenges of ATAC-seq.

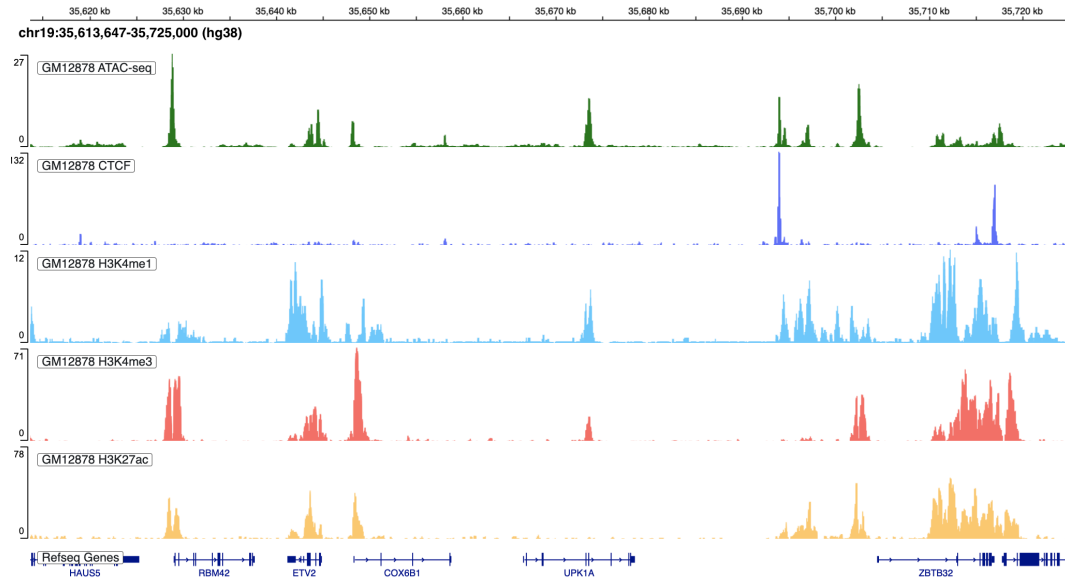


Figure 1.2: Signal coverage of ATAC-seq in comparison with other epigenome assays. The coverage profiles of the GM12878 cell line surrounding a representative locus, showing the multi-level epigenomic information ATAC-seq represents. Data are extracted from ENCODE and the genome view is supported by <https://igv.org/app/>.

1.3.1 Experimental Workflow of ATAC-seq

An ATAC-seq experiment can be seen as a variation to the standard NGS pipeline with modified pre-fragmentation. It uses a genetically engineered hyperactive *Tn5* transposase that simultaneously **cuts** in accessible genome regions randomly and **inserts** pre-integrated adaptors to these regions. During this process, the insertion sites are repaired with a 9-bp overhang. Enrichment is then performed by amplification of the inserted adaptor sequences, followed by NGS paired-end sequencing to generate high-throughput sequence fragments (Figure 1.3) [21–25]. The more ‘active’ (accessible) regions which are less occupied by nucleosomes and other proteins are more frequently digested through the transposase reaction that produces ATAC-seq **fragments**.

1.3.2 Standard Analysis of ATAC-seq

Similar to most other NGS assays, ATAC-seq data analysis includes *pre-processing*, *core analysis*, *advanced analysis*, and the potential *integration with multi-omics data* (Figure 1.4) [3]. Analysis generally starts from the raw sequencing data which are

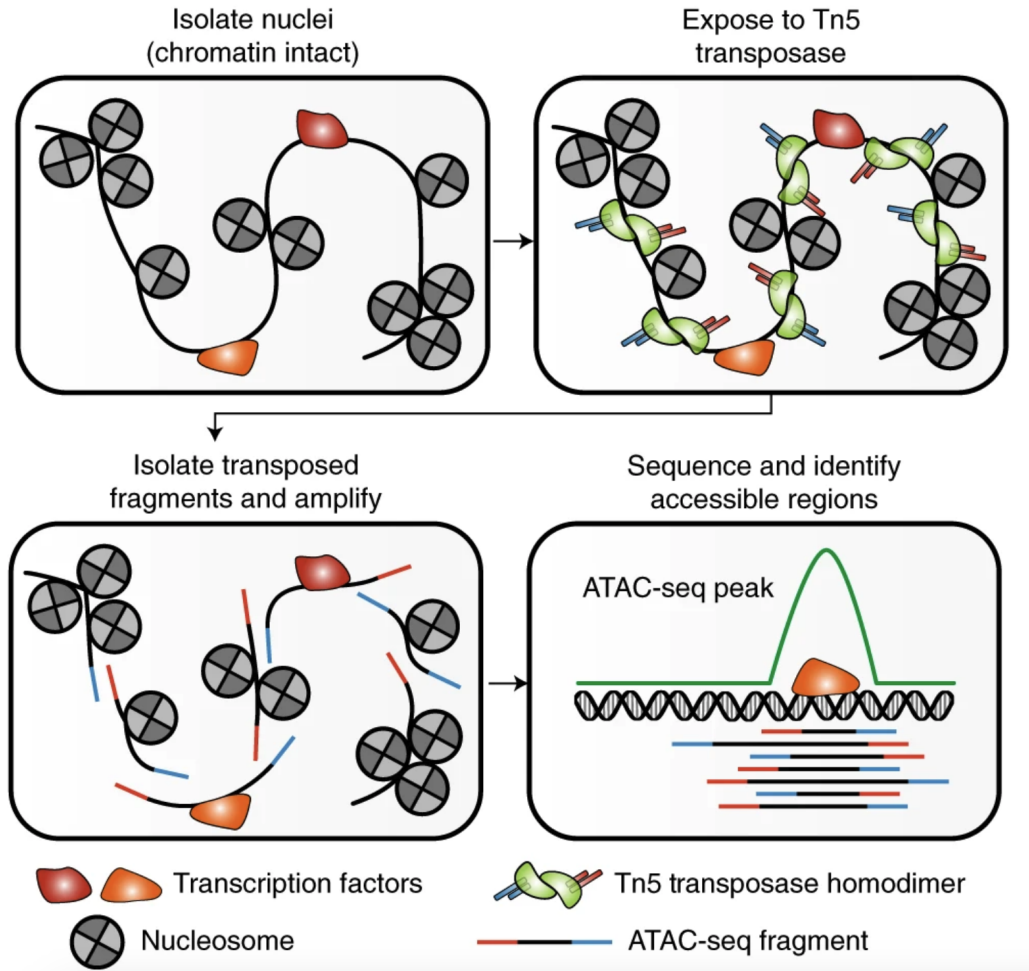


Figure 1.3: Overview of ATAC-seq experimental protocols. Figure from [25].

sequential reads of bases. *Pre-processing* including alignment (aligning reads to reference genome) and post-alignment processing (e.g. de-duplication and quality control) is similar across different assays ¹. *Core analysis* is often the most essential step to extract signals from background noise and serves the basis of downstream advanced analysis. In assays like ATAC-seq, DNase-seq, MNase-seq and ChIP-seq, the core analysis is **peak calling** (i.e. the identification of signal-enriched regions, referred to peaks). *Advanced analyses* in ATAC-seq include but are not limited to peak and motif enrichment, TF footprinting, and nucleosome positioning. *Multomics data integration* involves the integration with paired or un-paired assays and regulatory network reconstruction. *Quality Control (QC)* is carried out and

¹but may differ in actual algorithms, for example BOWTIE and BWA are used in alignment for ATAC-seq and ChIP-seq, while STAR is used in RNA-seq alignment for calling different transcripts

recorded throughout the pipeline to ensure the quality and confidence of each step. For raw sequencing data, one should check G+C content distribution, base calling quality and removing adaptors; During alignment, QC makes sure the retained reads are properly mapped and have Mapping Quality (MAPQ) scores higher than required threshold; For aligned reads, QC checks their size distribution² and TSS enrichment, removes PCR duplicates and mitochondrial-mapped reads; For ATAC-seq peak calls, it is a standard requirement to filter out regions overlapping with ENCODE blacklists³ [26] as they often produce false-positive enrichment problematic for functional genomics analysis [26].

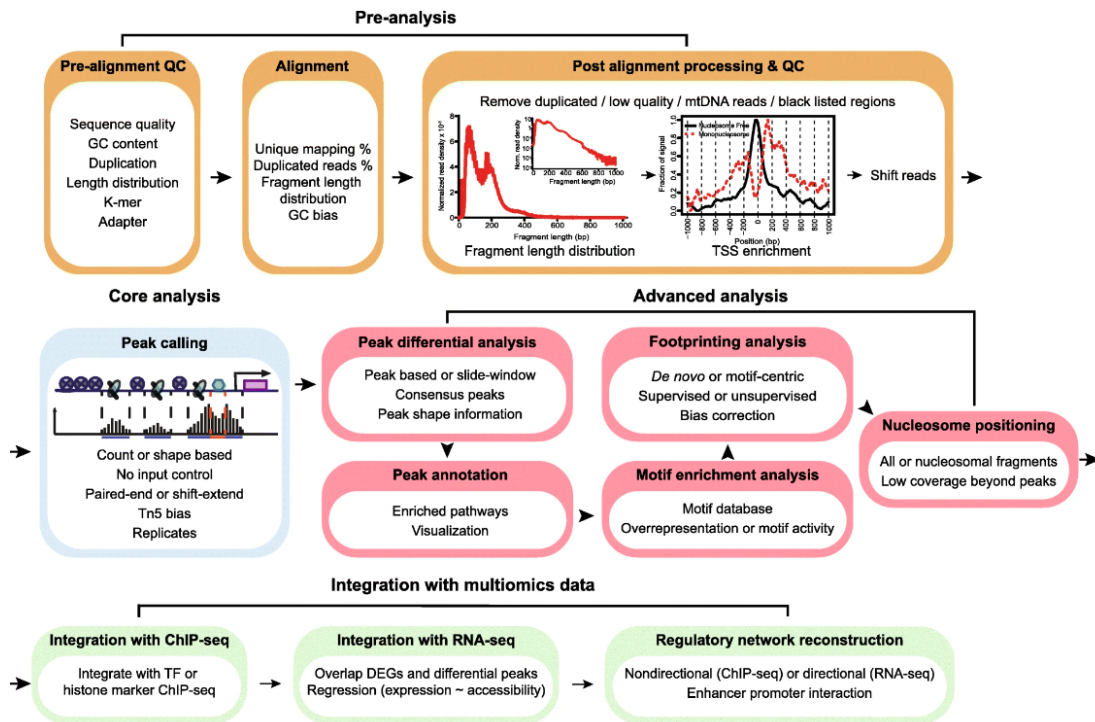


Figure 1.4: The standard analysis pipeline for bulk ATAC-seq data. The blue-boundary box highlights peak calling as the core analysis. The figure is adopted from [3].

²a successful ATAC-seq experiment should produce aligned reads of size distribution reflecting peak patterns at the lengths of mono-, di- and tri-nucleosomes

³anomalous, unstructured, or high signal in NGS experiments independent of cell line or experiment, due to difficulties such as collapsed repetitive regions in the reference sequence used in assembly

1.4 Single-cell ATAC-seq (scATAC-seq)

Section 1.3 talked about traditional **bulk** ATAC-seq assays, which takes cell population or tissue samples as input, reading out the aggregated or averaged signals of the mixture of genomics profile across cells. In fact, cell-to-cell heterogeneity are prevalent in biological tissues samples, such as peripheral bone marrow, neuronal progenitors and tumour micro-environment, but the signal of sub-grouped cells can be indistinguishable in the average profile, especially for functionally rare cell types. Recent emergence of single-cell sorting, labeling and sequencing technologies have pushed the resolution of genomics onto per-cell level and have rapidly applied to various functional genomics areas including single-cell RNA-seq (scRNA-seq), single-cell ATAC-seq (scATAC-seq), single-cell Hi-C (scHi-C) and more recently, integrative multi-omics assays, such as CITE-seq and paired scRNA + scATAC, which simultaneously profile multiple functional genomics types in parallel.

Note that the data quality of bulk ATAC and scATAC assays is inherently different. This is because scATAC-seq has a more sensitive experimental design, coupled with the availability to perform ‘cell calling’ to filter out a significant number of reads from non-cell barcodes. As shown in Figure 1.6C, scATAC-seq produces signals of higher quality and much less noise than bulk ATAC-seq, even when downsampling to the same sequencing depth. However, it is challenging to analyse and interpret scATAC-seq because of its extreme data sparsity (Section 1.4.2), especially in rare cell types whose biological signals are often hidden among the abundant cell types within the cell population.

In this section, I will introduce the mainstream experimental techniques and standard analysis workflow of scATAC-seq, providing the background information for the scATAC-seq analysis in Chapters 2 and 4.

1.4.1 scATAC-seq Experimental Approaches

The most essential step in single-cell sequencing is the isolation and identity-recognition of individual cells, which is realised through either physical sorting or single-cell barcoding approaches, or the combination of the two. The current

mainstream single-cell ATAC-seq platforms include plate-based combinatorial indexing (sciATAC and snATAC-seq) [27], plate-based microfluidics (Fluidigm C1 and Smart-Seq2) [28], and droplet-based GEM technologies (10X Genomics Chromium) [29]. Each platform relies on different barcoding strategy and/or media type.

Combinatorial indexing uses a two-step barcoding system to identify individual cells. The nuclei of lysed cells are first treated with barcoded transposases to digest into the first 96-well plates, then pooled before sorting by fluorescence-activated sorting (FACS) [30, 31] into the second 96-well plates when the second barcodes are introduced by PCR amplification. The cell identities are recognized by the unique combinations of two barcodes [27, 32]. By contrast, Fluidigm C1 [28] only applies a single-round barcoding which is during PCR. It utilizes a programmable microfluidics device to physically isolate single cells before performing transposase digestion and PCR on integrated fluidics circuit (IFC). These approaches are restricted by the nature of plate-based platforms that cannot process large number of cells in parallel. More recently, droplet-based 10X Chromium allows simultaneously profiling of massive number of cells (up to millions) [29, 33]. 10X Chromium developed Gel bead-in-Emulsion (GEM) to capsule individual cells with uniquely barcoded reagents, enabling cells in each droplet to be isolated, amplified and analyzed separately. Therefore droplet-based platforms support higher throughput, greater precision, and more automatic manipulation of sequencing at lower cost.

Above all scATAC-seq platforms, FACS is particularly useful to provide with individual cells *ground-truth* identities. It uses multi-parameter flow cytometry to separate and label cells based on known phenotypical features such as characteristic membrane proteins, which can be applied along with various scATAC-seq pipelines.

1.4.2 Standard Analysis of scATAC-seq

Most single-cell sequencing platforms were originally developed for scRNA-seq and have subsequently been adapted to scATAC-seq and other single-cell sequencing types. Employing any of the cell-sorting techniques as described in Section 1.4.1, the sequencing reads associated with a single cell are tagged with the same barcode.

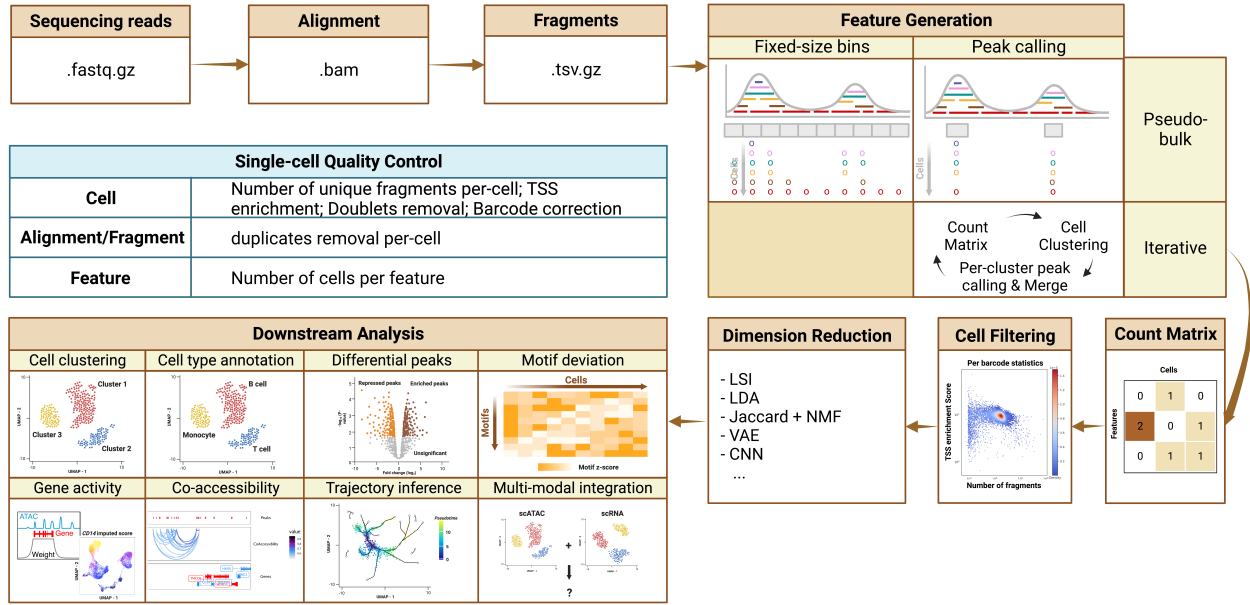


Figure 1.5: scATAC-seq Analysis Workflow. Created with BioRender.com.

Figure 1.5 shows a standard workflow of scATAC-seq analysis from raw sequencing reads from illumina platform to various downstream analysis, via *alignment* and *fragments creation*, *feature generation*, *count matrix generation*, *cell/barcode filtering*, and *dimension reduction*. All QC processes from bulk ATAC-seq (Section 1.2.2) are also applied for scATAC-seq, with additional QC on cell/barcode-specific criteria, including *barcode correction* and *nuclei-doublets removal*. From the raw region-by-cell count matrix, cells/barcodes with sufficient number of reads/fragments and TSS enrichment score above a minimum threshold are retained. Algorithms for alignment, read quality control and peak calling remain same as in bulk ATAC-seq, while in single-cell-specific steps there are a number of methods available based on different algorithms and logic, the combinations of which have also formed the ecosystem of consistent development of scATAC-seq data science pipelines over recent years. These pipelines either focus on addressing specific steps or provide end-to-end comprehensive analysis, each with their own advantages and limitations.

Prior to downstream analysis, there are three core steps involving the most variation of algorithms across pipelines: feature generation, count matrix generation and dimension reduction.

Feature generation identifies potential open chromatin regions from barcoded aligned reads. These open regions are then used as genetically informed *features* in downstream interpretation for the data. Current feature generation use either **peak calling** or genome-wide **fixed-size bins**. Peak calling in scATAC-seq, due to data sparsity, is impractical to perform on each single cell individually. Instead, it is performed on ‘pseudo-bulked’ scATAC-seq fragments aggregated across all cells to produce a common set of peaks. This approach is sub-optimal for the identification of peaks in rare cell types, the signal of which may easily be overwhelmed by signal and noise from more common cell types. Alternatively, one can define features as consecutive, non-overlapping fixed-size bins of genome segmentation, which provides a non-biased set of features and circumvents the potential to overlook of rare signals in peak calling [34–36]. However, the choice of the optimal bin size is subject to individual use cases, and analysis performing on the high number of bins is computationally extensive. Another approach, *iterative peak calling*, produces an initial set of peaks from pseudo-bulked data, which is used to construct a peak-by-cell count matrix to cluster cells, followed by peak calling on the aggregated alignments of data from each cell cluster before merging into a common ‘meta-peak’ set [37, 38]. This process can be performed iteratively to improve the sensitivity for calling cell-type-specific regions and enhancing signals from rare cell types which may be obscured in the original pseudo-bulked peaks, but it relies on a peak calling algorithm that can handle both high and low input read numbers. In special cases where FACS-labeled cell types are given, the FACS labels can be used instead of cluster identities with a single iteration [30, 39].

Count matrix generation uses the identified representative regions to create a region-by-cell count matrix, where the value of each entry C_{ij} is the count of the fragments or aligned reads overlapping with feature i in cell j . Theoretically, after the removal of PCR duplicates, a diploid genome has at most an accessibility count of two at each genomic position, leading to inherent data sparsity (1–15% of peaks

detected per cell [34, 40]). Various studies have found binarizing the count matrix ⁴ has better performance in downstream analysis irrespective of the particular method used [41], and binarization has become a common practice in most analysis pipelines.

Dimension reduction is often employed to reduce the data complexity of the high-dimensional region-by-cell count matrix to facilitate downstream analysis. Due to much higher drop-out rate than scRNA-seq, advanced statistical methods are required for low-dimensional embedding or modeling of scATAC-seq. As shown in Figure 1.5, existing algorithms including latent semantic indexing (LSI), Latent Dirichlet Allocation (LDA) [42, 43], Jaccard with Non-negative matrix factorisation (NMF), variational auto-encoders (VAE) [44] and convolutional neural networks (CNN), many of which are adopted from natural language processing. These methods above can be categorised as ‘sequence-free’ models that perform solely representation learning but ignore the information of actual DNA sequences of the regions [34, 45]. To infer correlations with TFs, it requires differential motif enrichment analysis, such as chromVAR [46] which employs a ‘sequence-dependent’ framework and scBasset [45] that extends the deep CNN model Basset [47] to scATAC-seq. I will discuss these related topics in Chapter 4 and 5.

1.5 ATAC-seq Peak Calling

As introduced in Section 1.3.2 and 1.4.2, peak calling is a crucial step in both bulk ATAC-seq and scATAC-seq analysis in order to identify candidate open chromatin regions for regulatory programs, serving as the basis for most downstream analysis and many predictive models for functional genomics [3]. For instance, in sequence-based machine learning models that predict gene expression and various epigenomic functions for candidate genomic regions (Section 1.7), peak calls are often used as input features, such as the per-segment binary accessibility of the genome. Therefore, the accuracy, precision, and robustness of peak calling are crucial for the reliability

⁴ $C'_{ij} = \begin{cases} 1, & \text{if } C_{ij} > 0 \\ 0, & \text{otherwise} \end{cases}$, where C represents the raw region-by-cell count matrix and C' is the binarized matrix

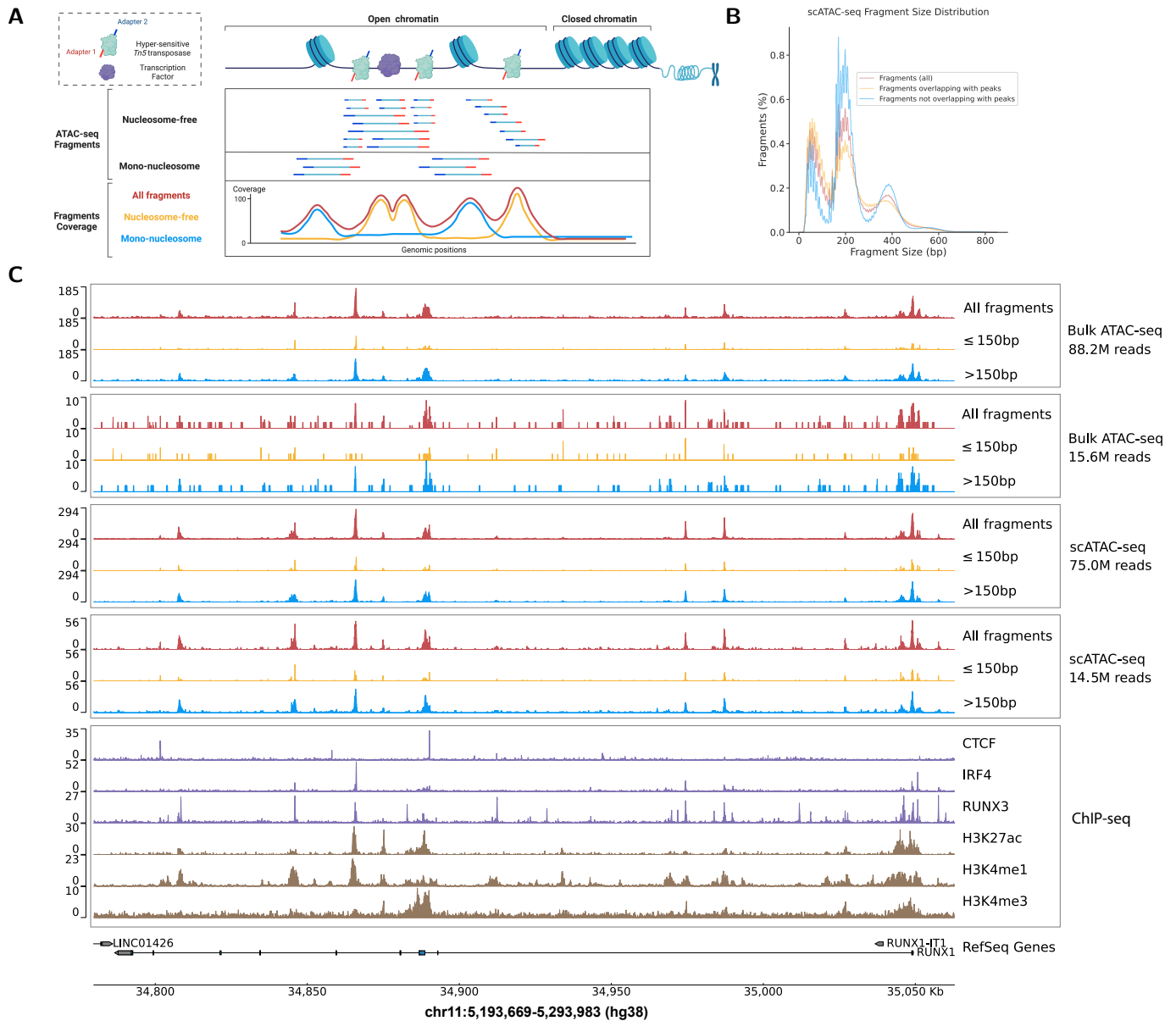


Figure 1.6: The fragment size of ATAC-seq data holds rich information of functional genomics. **A.** Schematic diagram showing *Tn5* digestion in competence with TFs and nucleosomes, resulting in different signal patterns in NFR and mono-nucleosome regions. Created with Biorender.com in reference with [3]. **B.** ATAC-seq fragment size distribution, grouped and colored by all fragments (red), fragments overlapping with peaks (yellow) and fragments not overlapping with peaks (blue). Fragments and peaks are taken from 10X scATAC-seq human GM12878 and mouse A20 mixture dataset, sub-setting only on GM12878 cells. **C.** GM12878 bulk and scATAC-seq datasets of different sequencing depth around the RUNX1 locus, grouped by all fragments (red), fragments ≤ 150 bp (yellow) and fragments ≥ 150 bp (blue). Reference ChIP-seq signals of TF binding (CTCF, IRF4 and RUNX3, purple) and Histone modifications (H3K27ac, H3K4me1 and H3K4me3, brown) are downloaded from ENCODE. Source of data: see Appendix 1.

and reproducibility for the model infrastructure. In this section, I will introduce the challenges of ATAC-seq peak calling and a brief overview of current algorithms.

1.5.1 Challenges of ATAC-seq Peak Calling

Accurate peak calling in ATAC-seq data is challenging due to the following factors:

1. The noisiness of the data. ATAC-seq has proven to be the most sensitive chromatin accessibility profiling technique to date. However, The *Tn5* transposase used in ATAC-seq has a preference for certain regions of the genome, leading to uneven coverage and noise in the data. The background noise can interfere with peak calling and result in false positives or false negatives. In many cases, it is unavailable to get abundant number of cells due to experimental limitations or the existence of rare cell types, which leads to low-quality ATAC-seq data with low signal-to-noise ratio. Therefore, the high degree of variability in the signal intensity across regions of open chromatin and the low-quality data bring difficulties to distinguish between true peaks and noise.

2. The lack of input control. Sequencing assays like ChIP-seq profiles fragments targeted by antibody as well as the control with no antibody used. ATAC-seq, on contrary, has no input controls available in which *Tn5* transposase inserts into protein-free DNA. As a result, a number of peak callers which require input controls are impractical for ATAC-seq.

3. The lack of distinct peak shapes. Unlike ChIP-seq, ATAC-seq lacks of a unified peak ‘shape’ due to the mixture of information it reflects. The fragments from ATAC-seq come from both NFR and nucleosome-bound regions (Figure 1.6A), bringing challenges to model the sequencing data based on a pre-defined statistical distribution.

1.5.2 Current ATAC-seq Peak Calling Algorithms

Peak calling for NGS data is fundamentally a statistical question. Traditional ATAC-seq peak callers build statistical models in two categories, shape-based and count-based [3]. Shape-based peak callers use read density profile and the distinct

shape of peak regions to improve peak calling in ChIP-seq [48], but has not yet been applied in ATAC-seq yet. Count-based peak callers compare read distribution shape in a candidate region to a random background using different assumptions on statistical models, such as Poisson distribution used in MACS2 [49] and zero-inflated negative binomial distribution employed by ZINBA [50]. They are generally easier to interpret and more widely used [3]. MACS2 is the most popular peak caller across assays and is the current ENCODE standard for ATAC-seq. Originally developed as a ChIP-seq peak caller, it uses a model-based approach to identify peaks by modeling the shift size of reads and using a Poisson distribution to estimate the significance of each peak. When being applied to paired-end ATAC-seq data, the option `-f BAMPE` defines fragments as the ends of the paired alignments with no modeling or artificial extension. The singleton (i.e. unpaired) alignments are ignored. MACS2 requires a large number of reads for peak calling, while different ATAC-seq datasets vary in read depth that it has been reported MACS2 may fail to identify true peaks or produce high number of false positives from insufficient read depth [3, 34]. F-seq [51] uses kernel density estimation to profile fragment distribution, and Genrich [52] is an unpublished peak caller that detects intervals centering on *Tn5* cut sites and provides analysis of multiple replicates.

Two recent peak callers have been designed specifically for ATAC-seq, HMMRATAC [53] and RCL [54]. HMMRATAC adopts a ‘decomposition and integration’ concept that first models layers of coverage signals corresponding to fragment sizes and then learn the cross-layer relationships using a three-state semi-supervised hidden Markov model (HMM), corresponding respectively to NFRs, nucleosome-binding regions, and background regions. Therefore apart from peak calling, HMMRATAC is also able to model and extract nucleosome positioning. However, in our experiments, HMMRATAC seems to produce valid peak calls only on high-quality datasets (see the benchmarking results of peak calling in Chapter 3 and 4) and requires extensive parameter adjustments for different datasets and is computationally intensive especially for large datasets ⁵. RCL (Replicative

⁵The developers recently reported HMMRATAC has been significantly accelerated and optimised in the MACS3 modules, but the evaluations are still underway.

Contrastive Learner) [54] specifically tackles the integration of multiple replicates for bulk ATAC-seq, using unsupervised contrastive learning to identify shared signals from the different replicates to optimise the signal extraction for reproducibility, but it has not adapted to cases including single-replicate datasets, heterogeneous cell population or scATAC-seq without known replicates.

LanceOtron [55] is a supervised deep-learning-based peak caller for ATAC-seq, ChIP-seq, and DNase-seq data. It combines a CNN module with local enrichment measurements to structure a combined peak scoring algorithm. For each candidate region, the CNN takes the base-pair resolution normalised signals of its central 2kb as input to address their signal abundance with respect to the relative positions and outputs a ‘shape score’. Along with the CNN module, a logistic regression model takes the enrichment measurements as input and outputs an ‘enrichment score’. The output of the two parts as well as 11 local ‘enrichment scores’ predicted from labeled peaks are combined into a multilayer perceptron to outputs a final ‘overall peak score’. Setting different cutoffs on the overall peak score, LanceOtron is able to produce peak calls at different significance levels. The supervised model architecture of LanceOtron requires high-quality labeled data for model training and can produce unsatisfactory performance for low-quality data.

Peak callers adopted from ChIP-seq or DNase-seq assume that ATAC-seq peak patterns share the same properties [3], and HMMRATAC learns the fragment size cutoffs for segmenting NFR, nucleosome-flanking and background signals also based on pre-defined statistical distributions. These assumptions may not be suitable for ATAC-seq signals, especially when considering the multi-level information ATAC-seq naturally represents. In addition, there is no well-developed benchmark framework as well as comprehensive evaluation criteria for ATAC-seq peak callers [3].

1.6 Multiple levels of Gene Regulation Represented in ATAC-seq Fragment Size

ATAC-seq fragment size is an important factor in functional analysis. It has been observed that fragments from NFRs tend to be shorter compared to those

from nucleosome-binding regions. Furthermore, TFs have been reported to bind more frequently in NFRs (Figure 1.6A). Hence, when analyzing the potential TF binding sites from ATAC-seq, it is reasonable to focus on shorter fragments (i.e. fragments shorter than 150 base pairs, approximately the size of a mono-nucleosome). However, this size filtering strategy is only valid given high quality data such as high-resolution scATAC-seq or bulk ATAC-seq aggregated from an abundant cell population, otherwise the short fragment extraction and subsequent loss of data can decrease the signal to noise in low-quality data (Figure 1.6C).

On the other hand, fragments of different sizes all contribute to ATAC-seq peaks, but peak regions have been observed to overlap with higher percentage of short fragments compared to non-peak regions (Figure 1.6B). Therefore, it is important to include fragments of both NFR and nucleosome-binding regions in the context of ATAC-seq peak calling, but different fragment sizes should be considered in different ways to obtain a high-resolution and bespoke analysis from the data.

1.7 Machine Learning Applications in Functional Genomics

The ever-growing data volumes and enhanced computational capacity over the last decade have made machine learning a valuable tool in many science domains. In genomics, data are often high-dimensional, complex and rich in non-linear relationships, which is difficult to analyse and generate plausible insights using statistical methods that are strongly dependent on pre-assumptions on features and domain expertise [56], especially given the fact that the systematic mechanism of gene regulation is still poorly understood. Therefore, machine learning, and particularly deep learning approaches, have been increasingly applied to genomics and generated powerful insights over conventional methods in tasks from data processing, unsupervised learning to functional predictions.

In this section, I summarise a subset of current machine learning applications in NGS-based functional genomics on different tasks, and specify the challenges when incorporating these approaches with genomics domain expertise.

1.7.1 Identifying and Annotating Functional Elements

Integrative analysis on existing large-scale, multi-cell-type sequencing data have identified and databased hundreds of consensus short sequences (~ 5 -20bp) known as ‘motifs’ specific to certain TFs (MEME [57], Cistrome [58], HOCOMOCO [59], JASPAR [60]), cell-type specific functional genome annotations (ChromHMM [61–64], SegWay [65], hiHMM [66]) and cell-type specific candidate *cis*-regulatory elements (cCREs) (SCREEN [7]). These integrated methods employ a variety of machine learning strategies. For example, ChromHMM uses a multivariate hidden Markov model (HMM) to infer the hidden states of pre-segmented fixed-size genomic intervals (e.g. 200bp) by explicitly modeling the combinatorial presence/absence from the observations of the epigenomic marks at per base-pair level on the basis of the significance of the observed count of sequencing reads relative to a Poisson background distribution. The learnt hidden states are then annotated as ‘chromatin states’ by calculating the most probable state for each interval, in which the chromatin states correspond to regulatory functionalities, including ‘active promoters’, ‘weak enhancers’, and ‘heterochromatin’ [63]. Since being proposed, ChromHMM has been used to annotate the genome into different number of states based on different range of epigenomic marks for more than 100 cell types respectively, as well as providing a universal genome annotation using a ‘stacked modeling’ approach by jointly profiling multiple cell types and assays [64].

Despite of all these efforts, the exact high-resolution functional decoding of the non-coding genome is far from completion. In addition, it can be unrealistic to segment the entire genome into mutually-exclusive functional states, such as at the boundaries of distant enhancer elements. All these challenges call for more advanced and systematic functional analysis based on the ever-growing NGS databases.

1.7.2 Signal Dnoise, Dropout Imputation and Batch Corrections

NGS analysis requires advanced statistical methodology to handle inherently missing data caused by technical imperfections, experimental operations or insufficient

sample size. In practical uses, signal denoise and dropout imputation are continuous challenges in data preprocessing.

Single-cell data is another type of data that suffers dropout events, leading to missing values that obscure the true observations in the original sparse matrices. For scRNA-seq data, several imputation methods have been developed to estimate the true expression values for genes by leveraging information from the neighbourhood cells. Different algorithms represent the cell-to-cell similarity in various ways. MAGIC (Markov Affinity-based Graph Imputation of Cells) [67] uses diffusion geometry to learn the inherent manifold of the cells from the k-nearest-neighbour cell-to-cell distances, in which the inferred manifold serves as imputation. SAVER (Single-Cell Analysis Via Expression Recovery) [68] uses a VAE structure that imputes missing values by sampling from the conditional distribution of the missing expression values based on the generative model on the observed data. scGNN [69, 70] leverages graph neural networks where cells are represented as nodes and the edges capture cell-to-cell relationships representing expression similarity.

1.7.3 Functional Predictions

Predicting biological functions solely from the genome sequences is a fundamental and challenging goal in genomics. Currently, deep learning models are exceptionally successful in these tasks in terms of prediction accuracy, feature extraction and the ability to generalize. They produce single or multi-class predictions for functional genomics ranging from TF binding, chromatin accessibility, gene expressions and 3D chromatin interactions. The current popular architectures include CNNs and their variations, used in DeepBind [71], DeepSEA [72], Basset [47], and Basenji [73], etc., and more recently the application of attention modules in transformers, employed by Enformer [74], etc.

1.8 Aims and Scope of the Thesis

The overall goal of this thesis is to demonstrate the different uses of ATAC-seq fragment size for different analytical purposes, specifically, i) apply a simple fragment

size filtering strategy to construct a high-resolution scATAC-seq analysis pipeline to prioritise functional SNPs; and ii) developing an ATAC-seq specific peak caller as a mechanistic algorithm integrating both signal coverage and fragment size, to achieve high resolution, high specificity and robustness to low-quality ATAC-seq and scATAC-seq data. In addition, I will explore the use of iterative peak calling in improving differential motif enrichment analysis in scATAC-seq data.

The sections above in Chapter 1 illustrate the background information of the non-coding genome, NGS assays for functional genomics, and the challenges in current NGS analysis which call for improvement of statistical modeling and biological interpretation. These motivate the study in the rest of the thesis, summarised as follows.

Chapter 2 introduces AVOCATO, a collaborative work developing an end-to-end scATAC-seq pipeline for high-resolution functional analysis accompanied by a powerful interactive graphic interface for the visualisation and manipulation of processed data. As explained in Section 1.6, AVOCATO employs a simple but useful fragment size filtering strategy on high-quality scATAC-seq data to prioritise functional SNP variants, which I demonstrate using an example dataset. In addition, I explain in detail the optimised statistical algorithms in the pipeline.

Chapter 3 introduces EpiCall, a novel high-resolution peak caller designed exclusively for ATAC-seq. As explained in Section 1.6, ATAC-seq fragments of small and large sizes all contribute to peaks but not equally, thus they should be considered differently. This drives the development of EpiCall as a mechanistic model for the actual formation of sequencing fragments, making optimal use of fragment size and coverage information to identify peak regions. To demonstrate the comprehensive evaluation of EpiCall, I show the performance of EpiCall against current peak callers MACS2, HMMRATAC and LanceOtron across several public datasets in reference with the cell-type specific chromatin state annotations, TF motif enrichment, ChIP-seq signals of TF binding and histone modifications and evolutionary conservation. I also evaluate EpiCall in read-downsampled data of

low sequencing depth, demonstrating the robustness of EpiCall in peak calling data with sub-optimal quality.

Chapter 4 extends the use of EpiCall from bulk ATAC-seq to scATAC-seq data. Similar to Chapter 3, I show the robustness of EpiCall in peak calling cell-downsampled scATAC-seq data. Then I present the application of EpiCall in analysing of differential TF motif enrichment across single cells and the clusters of cells, along with the exploration of a single-step iterative peak calling strategy to facilitate the analysis. In addition, I demonstrate the ability of EpiCall in detecting and distinguishing nearby peaks sensitively, a superior characteristic in comparison with other peak callers.

Chapter 5 concludes the thesis and envisions the potential future directions of the field.

2

AVOCATO: A High-resolution Analysis Pipeline for scATAC-seq Data

Contents

2.1	Introduction	26
2.2	Results	27
2.2.1	The Characteristics of AVOCATO	27
2.2.2	The Functionalities of the Workflow	28
2.2.3	Configuration, Execution and Visualisation	30
2.2.4	Example: Visualisation of 10X PBMC 3K scATAC-seq data analysis	31
2.3	Methods	33
2.3.1	Whitelist-free barcode correction	33
2.3.2	Adapting ArchR for Stage 1 Analysis	35
2.3.3	Iterative Decision Tree for Selecting Cluster-Representative Genes	36
2.4	Code Availability	36
2.5	Acknowledgement	36

2.1 Introduction

The challenges in handling scATAC-seq lie in the high dimensionality and sparsity of data, as well as the biological heterogeneity and experimental imperfection. Practically, analyzing scATAC-seq involves multi-stage data processing including

bespoke quality control, signal detection, and functional interpretation. These requirements have driven the continuous development of pipelines and tools focusing either on specific steps or providing end-to-end analysis from raw data to downstream interpretations. Although several tools addressing these issues have been developed, there is room for improvement, particularly with regard to higher resolution and sensitivity in signal processing. In this chapter, I will introduce a new scATAC-seq analytical platform, **AVOCATO** (**A**nalysis and **V**isualization **O**f single-Cell **A**TAC-seq **O**bservations), which was developed to produce high-resolution functional genetics identification and interpretation from scATAC-seq data.

As introduced in Chapter 1, AVOCATO adopts a simple but useful fragment size filtering approach for high-quality scATAC-seq data to retain signals from fragments $\leq 150\text{bp}$ only, which is specifically suitable for prioritising TF-binding related functional SNP variants from high-quality scATAC-seq data. In this chapter, I will walk through the pipeline and use the 10X PBMC 3k dataset to demonstrate the analytical results along with the user-friendly interactive visualisation platform.

2.2 Results

2.2.1 The Characteristics of AVOCATO

Instead of constructing a comprehensive pipeline integrating the state-of-the-art established processing methods, we design AVOCATO to be an application focused, scalable pipeline focusing on high-resolution, cluster-specific functional analysis. AVOCATO is built to provide the following characteristics:

- 1 **An End-to-end pipeline supporting multiple entry points.** We aim to allow both FASTQ and BAM formats as inputs, followed by three stages of analysis conducting step-wise quality control, dimensionality reduction and cell clustering, gene score activity inference, and cluster-specific functional analysis. Section 2.2.2 introduces the detailed functionalities of each stage.

- 2 **Graphic interfaces for processing and interactive visualizations.** Along with the analyses, we have provided AVOCATO with graphic interfaces

to set up program parameters, execute commands and visualize results. We use FLASK for constructing the backbone of the interfaces assisting with configuration, program execution, and the display of static results. We also provide interactive analysis of the high-dimensional results and customised inter-stage connections using Multi-Dimensional Viewer (MDV) [75] and Multi-Locus Viewer (MLV) [76], two powerful and efficient JavaScript tools designed for visualising multi-dimensional genomics data of varying types. Section 2.2.3 introduces the functionalities of the user interface and Section 2.2.4 demonstrates the use case on a public dataset.

3 High-resolution functional analysis facilitating the use of scATAC-seq in population genetics. In Stage 2, AVOCATO provides intersects analysis between scATAC-seq signals and the SNP variants derived from GWAS databases. We notice that retaining only short fragments $\leq 150\text{bp}$ produces high-resolution scATAC-seq signals which provides prioritisation from a subset of SNPs on functional regulatory elements such as enhancers and promoters which are uncovered within the raw scATAC-seq signals.

4 Reproducibility and scalability. AVOCATO is implemented using Snakemake [77] and follows its logic of pipeline construction and execution. The main functionalities are written in Python, R and bash. Each functionality is specified as a Snakemake ‘rule’ within ‘smk’ files. To execute a specific rule, simply specify the rule name `snakemake RULENAME -config=PATH/TO/CONFIG.yaml` or the output file `snakemake OUTPUT -config=PATH/TO/CONFIG.yaml`, Snakemake will automatically figure out the command dependencies and run all required steps. Therefore, the results generated by different configurations are tractable and reproducible, ensuring the scalability of the tool.

2.2.2 The Functionalities of the Workflow

We design the pipeline of AVOCATO (Figure 2.1) into three sequential stages, with each the following functionalities:

Stage 0: Pre-processing

Stage 0 pre-processes the raw scATAC-seq data into a fragments file (.tsv.gz). For the input, AVOCATO accepts paired-end scATAC-seq data in either FASTQ or BAM formats. For FASTQ input, AVOCATO requires all files are named in 10X-style format and stored in a single directory. AVOCATO provides the choice between Bowtie2 and BWA-MEM2 for read alignment, followed by a whitelist-free barcode correction step to fix the possibly incorrect cell barcodes (Details referred to Methods 2.3.1). After retaining the reads which are properly aligned and of high mapping quality, AVOCATO calculates the per-barcode read statistics and per-read barcode statistics for marking read duplicates and barcode filtering. The final BAM is then converted into a 10X-style fragments file, in which each line marks the start and end genomic position of a fragment, its corresponding cell barcode and the number of PCR duplicates. For BAM input, users should specify in configuration whether it requires quality control such as barcode correction and cell filtering.

Stage 1: Clustering and Marker Gene Selection

Starting from the fragments file, AVOCATO employs ArchR [36] for fixed-size-window based dimensionality reduction, cell clustering, gene activity calculations and imputations. Based on the imputed gene vs. cell score matrix, AVOCATO uses an iterative Decision Tree approach (Methods 2.3.3) to select the most representative down-regulated and up-regulated genes for each annotated cell cluster. The final output of stage 1 is a comprehensive table recording for each cell the clustering annotations, UMAP (Uniform Manifold Approximation and Projection) embeddings, activity scores of representative genes and QC metrics. An compressed H5 file is automatically generated for loading into MDV (Figure 2.3).

Stage 2: High-resolution Analysis and Variant Prioritisation

Based on the clustering annotations from the Stage 1 output, AVOCATO calculates per-cluster coverage tracks and performs peak calling giving the pseudo-bulk fragments. In this step, AVOCATO allows the user to choose whether to filter the

fragments according to their sizes, as in our experiments retaining the fragments with size less than 150bp improves high-resolution functional annotations. The peak calls are then used to intersect with provided genetic changes, such as the SNP variants from GWAS databases, to calculate statistical enrichment ([78, 79]), from which the most enriched clusters will be prioritised for downstream analysis. Therefore, the high-resolution analysis in Stage 2 can prioritise variants informative for TF binding in regions with most significant evidence in scATAC-seq.

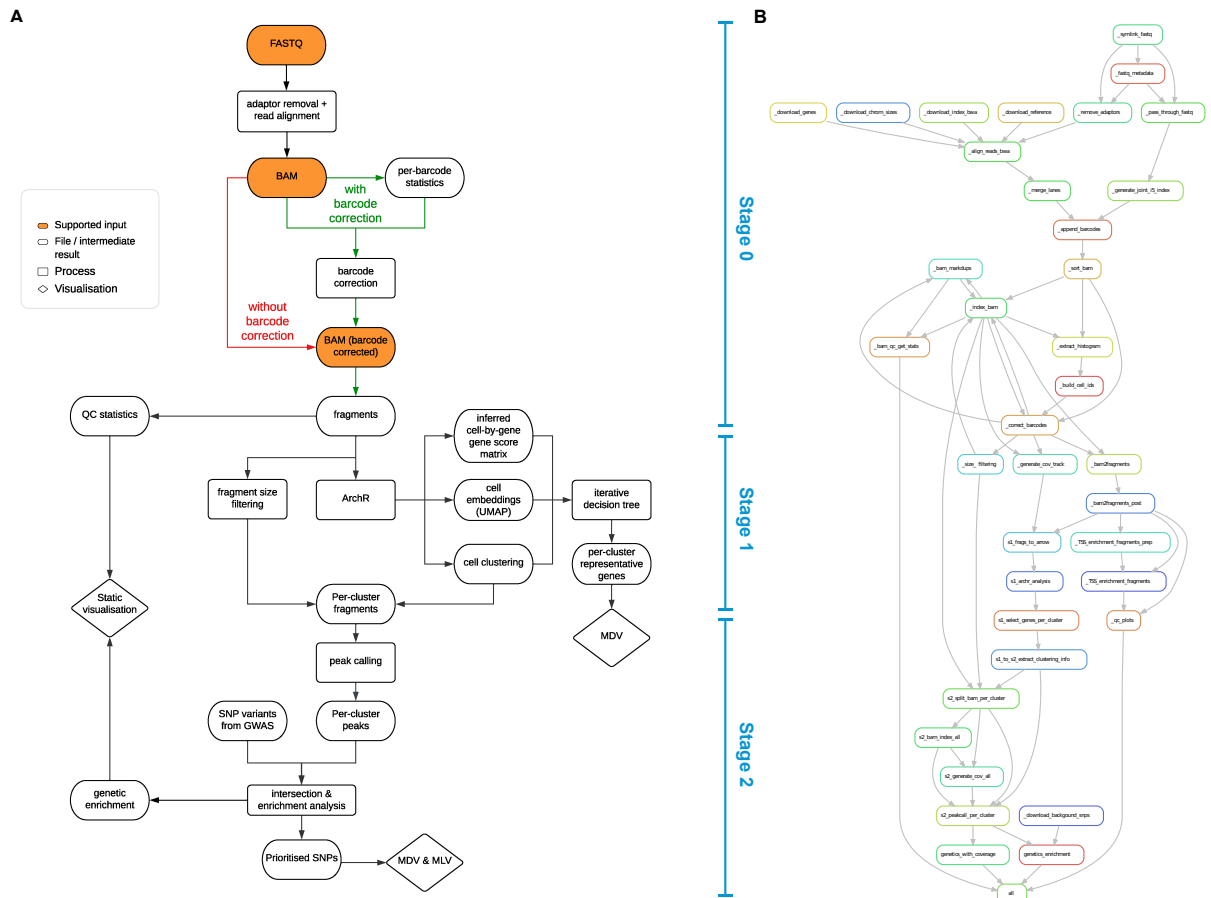


Figure 2.1: The End-to-end Workflow of AVOCATO. **A.** The functionalities and workflow of AVOCATO illustrated in a flowchart. Created by <https://lucid.app/>. **B.** The actual snakemake workflow of AVOCATO generated with `snakemake -dag`, in which the text in each box is the rule name. The rough separations of Stage 0, 1 and 2 are marked in the middle.

2.2.3 Configuration, Execution and Visualisation

We provide a template configuration file `config/config.yaml` containing all parameters and their default values required for AVOCATO. Users should modify

the configuration for individual use cases, or use AVOCATO’s interactive platform to easily customise the parameters and check for validation (Figure 2.2B).

As introduced in Section 2.2.1, the functionalities of AVOCATO are constructed using Snakemake ‘rules’. for example, the rule `all` runs end-to-end throughout the entire pipeline, while rules `all_stage1` and `all_stage2` runs each stage alone. Apart from the command line, AVOCATO offers a graphical interface to support interactive execution and data visualisation. (Figure 2.2B).

Apart from the configuration helper and pipeline running, the main use of AVOCATO’s interactive platform is visualization. From the page displaying the full list of previous analyses (Figure 2.2C), one can access to each project’s main page where the QC and static results of each stage are available. Along with the static displays, the core of visualisation is the dynamic, interactive platform powered by MDV, in which users can have full understanding of the entire dataset as well as each cluster or hand-picked cells with the facilitation of a variety of plot types, enabling customised exploration of the complex, high-dimensional scATAC-seq analysis results and potentially, utilising visualisation to guide downstream analysis. In the next section, I will demonstrate AVOCATO’s visualisation platform using 10X PBMC 3K scATAC-seq data as an example.

2.2.4 Example: Visualisation of 10X PBMC 3K scATAC-seq data analysis

I used AVOCATO to process 10X PBMC 3K scATAC-seq data end-to-end from the raw FASTQ files. Figure 2.3 demonstrates using the AVOCATO interactive platform to visualise results of each stage.

For Stage 1, AVOCATO displays quality control statistics as static plots including per-barcode mean TSS enrichment vs.fragment count (scatter plot), fragment size distribution (histogram), TSS enrichment score distribution (line plot) and per-barcode fragment count distribution (histogram) (Figure 2.3A). The MDV interactive platform takes the output of Stage 1 and automatically generates interactive plots including 2D and 3D UMAP embeddings with annotated clusters

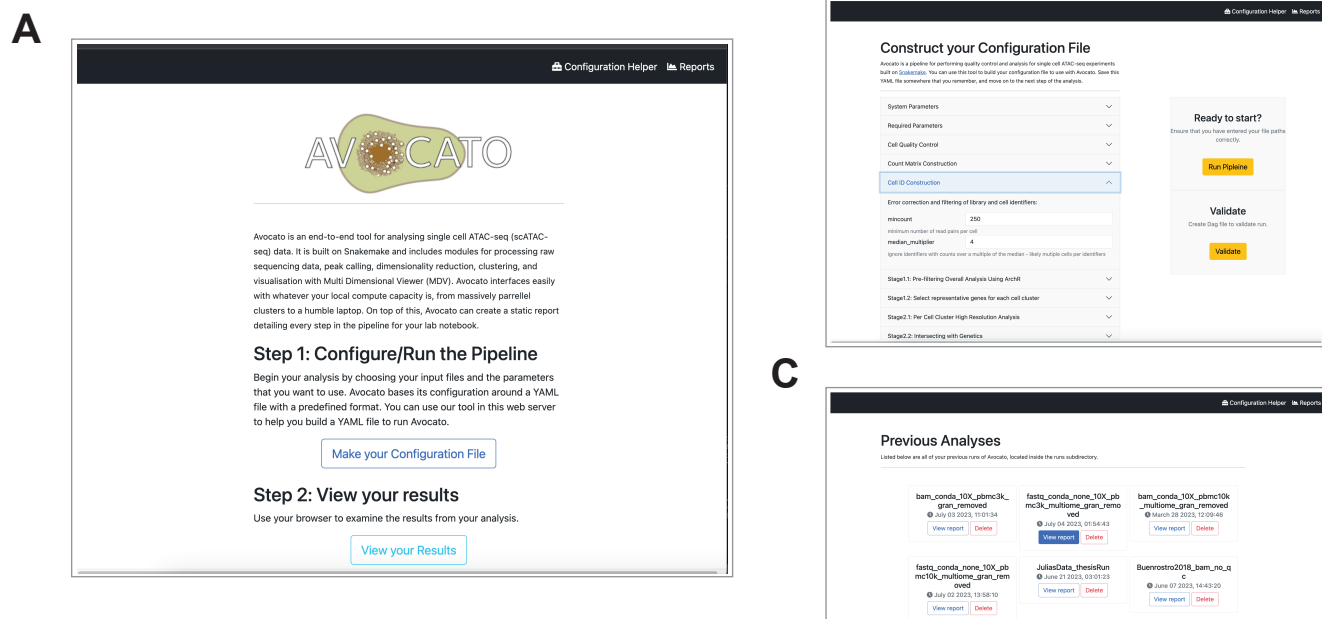


Figure 2.2: The graphical user interface of AVOCATO. The base page (A) guides the two steps of running AVOCATO. Clicking ‘Make your Configuration File’ in Step 1 directs to the configuration helper page (B) to set up parameters for creating the Snakemake config file; Clicking ‘View your Results’ in Step 2 directs to the results displaying page (C) which lists all previous analyses.

(Figure 2.3B left panel, top-left and bottom-left), per-cluster imputed gene scores (Figure 2.3B left panel, top-right) clustering information, number of cells per-cluster (Figure 2.3B left panel, bottom-right), a table listing all gene annotations of the corresponding genome build (Figure 2.3B right panel, top) and per-cluster fragment coverage tracks at specific genomic regions (Figure 2.3B right panel, bottom). Notice that all the plots are dynamic and cross-interactive. For example, in the top table on the right panel, selecting a gene automatically zooms the bottom coverage tracks onto the gene locus, as shown in Figure 2.3B in which the locus of B-lymphocyte marker gene *MS4A1* is cross-selected on both. Similarly, selecting a specific cluster(s) in any plot highlights the cluster-specific information in all other plots (Figure 2.3C). MDV also supports free-selecting any group of cells instead of cluster annotations. In addition to default plots, MDV has a wide range of available plots for different user cases and (Figure 2.3B, right panel, top-right toggle box) which all support customisation and interactions.

For Stage 2, AVOCATO conducts cluster-specific peak calling and statistical enrichment with GWAS associations. The SNPs can be assessed interactively with different levels of scATAC-seq signal abundance (i.e. fragment counts). Since the example dataset is PBMC cells of the immune and blood system, I chose variants associated with an immune-system related trait, Leukocyte Diseases (LD), for a reasonable validation analysis. Specifically, I used lead SNPs from the NHGRI-EBI GWAS Catalog [12, 13] and genotype-imputed SNPs from the a hematological trait GWAS analysis [80]. The ideal result will prioritise more significant enrichment in leukocyte clusters than the rest of the cells. The static displays of Stage 2 show the distribution of the number of peaks intersecting with SNPs (Figure 2.3D, top), in which different groups on the x -axis represent peak calls generated from fragments with or without size filtering (i.e. entire data, $\leq 150\text{bp}$ and $> 150\text{bp}$). The per-cluster statistical enrichment of peak-SNP intersections (Figure 2.3D, bottom) identify Cluster 8 and 5 with high enrichment while the rest remaining low, suggesting the two clusters consist of candidate cells to prioritise LD-associated traits. The results are consistent with the cluster identities inferred from the marker genes of the paired scRNA-seq data (B lymphocyte cells), indicating AVOCATO high-resolution cluster-specific analysis is valid for genetic variant prioritisation.

2.3 Methods

2.3.1 Whitelist-free barcode correction

In droplet-based scATAC-seq platforms, such as 10X scATAC-seq, the barcode used to label each single droplet and thus distinguish single cells are 10–20 base oligonucleotides within specific reads. In the original 10X Cell Ranger ATAC barcode correction protocol, a whitelist of valid barcodes are given in advance as a reference, and all observed barcodes that are not in the whitelist are then corrected by finding all valid barcodes within one mismatch from the observed sequence, and scoring them based on the abundance of that barcode in the read data and quality value of the incorrect bases, with an observed invalid barcode

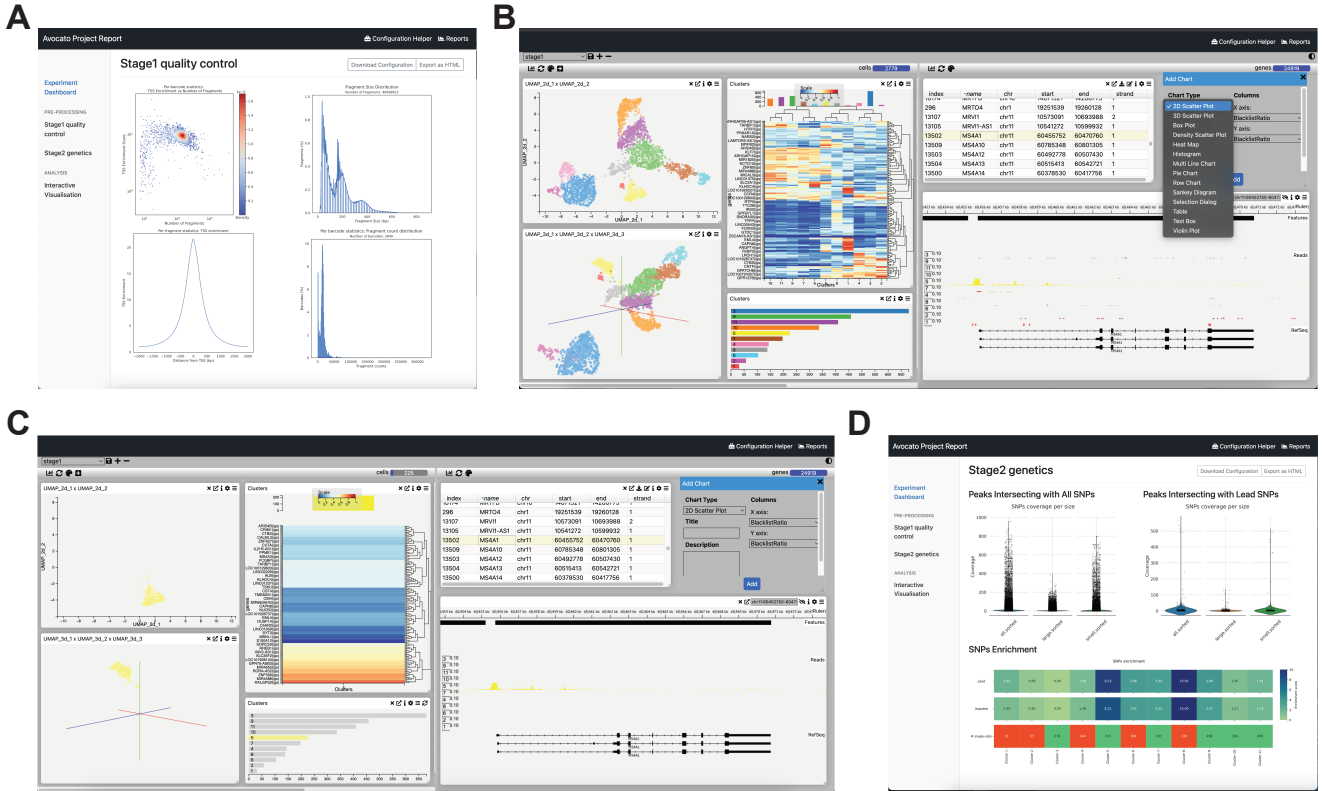


Figure 2.3: The visualisation platform of AVOCATO demonstrating 10X PBMC3k scATAC-seq data analysis.

being corrected if it has over 90% probability of being valid ¹. This approach relies on a whitelist which is not necessarily available in all cases, and is not able to correct barcodes with more than one mismatch.

Instead, AVOCATO implements a whitelist-free barcode correction algorithm which uses the observed barcodes only. From a histogram of observed raw barcodes we identify valid barcodes with high read counts, low read counts at Hamming distance 1, and low expected contribution from any other detected barcode (with possibly much higher frequency) at higher Hamming distances. Using these whitelisted barcodes, nearest-neighbour error correction is used to correct errors, and corrections are accepted if a unique near neighbour exist at Hamming distance at most 1. (Algorithm 1).

¹<https://support.10xgenomics.com/single-cell-atac/software/pipelines/latest/algorithms/overviewbarcode>

Algorithm 1: Whitelist-free correction for droplet-based oligonucleotide sequencing barcodes

Input: M - list $[(c_1, b_1), (c_2, b_2), \dots, (c_n, b_n)]$ of barcodes b_i and counts c_i , reverse sorted on c_i ; L - length of each barcode (same across all cells); $maxBaseError := 0.025$, $maxContamination := 0.02$, $minFracOfAverage := 0.02$

Output: M_{white} - barcode whitelist

```

1 Initialize  $TotCount = 0$ ;  $MeanCount = 1$ ;  $M_{white} := []$ 
2 for  $c_i, b_i \in M$  do
3    $minCount := minFracOfAverage \times MeanCount$ 
4   while  $c_i \geq minCount$  do
5      $maxD1 = c_i \times maxBaseError \times L$ 
6      $maxD2 = c_i \times maxBaseError^2 \times L(L-1)/2$ 
7      $D1 = \sum_j c_j$ ,  $\forall c_j, b_j \in M$  with  $Hamming(b_j, b_i) = 1$ 
8     if  $D1 < maxD1$  then
9        $D2 = \sum_j c_j$ ,  $\forall c_j, b_j \in M$  with  $Hamming(b_j, b_i) = 2$ 
10       $influence = maxBaseError^2 \times D2$ 
11      if  $influence \leq c_i \times maxCont$  then
12        if  $D2 < maxD2_i$  then
13          | Assign barcodes at Hamming distance 1 from  $c_i$  to  $c_i$ 
14        end
15        if  $c_i \geq minCount$  then
16          | Add  $b_i$  to  $M_{white}$ 
17          |  $TotCount += c_i$ 
18          |  $MeanCount = TotCount / |M_{white}|$ 
19        end
20      end
21    end
22  end
23 end
24 return  $M_{corr}$ 

```

2.3.2 Adapting ArchR for Stage 1 Analysis

AVOCATO involves ArchR in the Stage1 analysis. Starting from the fragments file, ArchR is used to perform dimensionality reduction (iterative Latent Semantic Indexing, iLSI), cell clustering (Seurat), single-cell embedding (UMAP), inferring gene score activity matrix and its imputations. In particular, ArchR employs an efficient storing strategy to create an HDF5-format ‘Arrow File’ stored on disk, allowing rapid access in R and the creation of a *region vs. cell* ‘Tile matrix’ of fixed-size windows as small as 500bp, which is used for the default dimensionality

reduction as well as the downstream analysis. We use the default cell clustering approach in ArchR, implemented by Seurat [29], which applies a graph-based clustering approach to annotate groups of cells based on the k -nearest neighbors (k -NN) cell-to-cell graphs. We use UMAP for projecting cells onto 2-dimensional and 3-dimensional embedding spaces for visualisation. At last, ArchR derives a *gene vs. cell* gene scores activity matrix which is distance-based, gene-size-bias addressed and depth normalized [36], followed by using MAGIC to impute gene scores by smoothing signal across neighbouring cells [67]. The final imputed gene score matrix is used to select a set of representative genes in the next step.

2.3.3 Iterative Decision Tree for Selecting Cluster-Representative Genes

To select a gene set that distinguishes all cell clusters, we designed an algorithm to identify representative genes for each cluster relative to all others based on the ArchR inferred and imputed per-cell gene score matrix from the scATAC-seq data. In addition to over-expressing genes which are normally considered as marker genes, we also include under-expressed genes, which in certain cases also help to uniquely distinguish clusters. The algorithm uses an Iterative Decision Tree (ItDT) approach in which for each cluster n a one-vs-others Decision Tree Classifier is grown to classify the cells into $cluster_n$ and $\neg cluster_n$, from which we identify the features of all non-leaf nodes to be the representative genes of $cluster_n$ (Algorithm 2).

2.4 Code Availability

The code of AVOCATO is available at

<https://github.com/Genome-Function-Initiative-Oxford/AVOCATO> with installation guide, documentation and tutorials.

2.5 Acknowledgement

AVOCATO is a collaboration project within Weatherall Institute of Molecular Medicine, University of Oxford, thanks to the contributions of Christopher Cole,

Algorithm 2: Iterative Decision Tree algorithm for representative gene selection

Input: X - the imputed gene score matrix, in which X_{ij} is the imputed gene score of gene i in cell j ; N - the number of cell clusters; G_{max} - the maximum number of representative genes to be selected for each cluster; AUC_{min} - the minimum AUROC for terminating the iterations; r - the train/test split ratio

Output: *GeneSet*

```

1 for cluster  $n \in [1, N]$  do
2   Randomly split  $X$  into train and test set  $X_{train}$ ,  $X_{test}$  at ratio  $r$ ,
   stratified for cluster  $n$ 
3   Initialize  $AUC_{train,n} \leftarrow 1$ ,  $AUC_{test,n} \leftarrow 1$ ,  $G_n \leftarrow 0$ ,  $GeneSet_n \leftarrow \{\}$ 
4   while  $AUC_{train,n} \geq AUC_{min}$  &  $AUC_{test,n} \geq AUC_{min}$  &  $G_n \leq G_{max}$  do
5     Grow a Decision Tree classifier  $DT$  given  $X_{train,n}$  to assign samples
     to  $cluster_n$  and  $\neg cluster_n$  (one-vs-others)
6     for gene  $g \in \{features\ of\ all\ non-leaf\ nodes\}$  do
7       Append  $g$  to  $GeneSet_n$ 
8        $G_n \leftarrow G_n + 1$ 
9       Remove gene  $g$  from  $X_{train}$ ,  $X_{test}$ 
10    end
11    Update  $AUC_{train,n}$ ,  $AUC_{test,n}$ 
12  end
13 end
14 return  $GeneSet$ 

```

Martin Sergeant, Emine Ravza Gur, Simone Riva and the supervisions from Jim Hughes and Gerton Lunter.

My independent contributions towards AVOCATO involve constructing the Snakemake pipeline and implementing the majority of functional scripts for Stage 0 (Setting up different entry points of the pipeline, quality control, and post-processing for BAM and fragments file), the entire Stage 1, and parts of Stage 2 (scripting and parallelizing cluster-specific BAM extraction, coverage calculation and peak calling). My collaborative contributions involve designing and implementing the barcode correction step (Methods 2.3.1), setting up different read aligners with G.L. and C.C., and creating the static data visualisation platform with C.C.. The rest of AVOCATO are developed by mixed collaborative work, in which the general contributions are as follows: E.R.G. and S.R. performed analysis for functional analysis; S.R. optimised the overall Snakemake pipeline and implemented the

algorithm for the genetic intersection and enrichment analysis in Stage 2; C.C. set up the initial pipeline structure, implemented the basic form of Stage 0 and wrote the first-version documentation; M.S. developed MDV and integrated it into AVOCATO; G.L. and J.H. proposed the project and provided supervision throughout the development. Stephen Taylor provided useful discussions and suggestions at the initial phase of the project.

3

EpiCall: A High-Resolution Peak Caller for ATAC-seq Data

Contents

3.1	Introduction	40
3.2	EpiCall	43
3.2.1	The Model Concept	43
3.2.2	Derivation of the Probabilistic Model	43
3.2.3	Implementation With hidden Markov Model	47
3.2.4	Calling Peaks With the Viterbi Algorithm	48
3.2.5	Defining Peak Score	48
3.3	Results	48
3.3.1	Accurate and sensitive detection of cell type-specific open chromatin regions	49
3.3.2	EpiCall sensitively recovers TF binding and functional histone modification signals	53
3.3.3	High Evolutionary Conservation in EpiCall Peaks	59
3.3.4	Robustness to low sequencing depth	61
3.3.5	Integrating fragment size and coverage benefits peak calling	62
3.4	Discussion	66
3.4.1	High-resolution peak calls enable accurate signal prioritisation of epigenomic profiling	66
3.4.2	Future work	66
3.5	Code and Data Availability	69
3.6	Methods	69
3.6.1	Data Preprocessing	69
3.6.2	Peak Calling	69
3.6.3	AUPRC measurement of ATAC-seq peak calls	71
3.6.4	Motif enrichment analysis	72

3.7 Acknowledgement 72

3.1 Introduction

In Eukaryotic cells, genomes are compacted into nucleosomes to different degrees resulting in a gradient of chromatin accessibility, ranging from condensed, inactive heterochromatin to nucleosome-free, active euchromatin. The accessibility of chromatin is positively associated with the activity of gene expression and *cis*-regulatory elements such as promoters and enhancers, probably in part because regulatory proteins such as transcription factors (TF) physically compete with nucleosomes for DNA binding. Chromatin accessibility is highly dynamic and heterogeneous between different cell types and tissues; cells also undergo systematic changes in chromatin accessibility during development and progression. Thus, precise characterization of the genome-wide chromatin accessibility landscape is vital for understanding cell and tissue-specific gene regulation and can give insight into the molecular mechanisms of cell development and disease-associated genetic mutants.

Among existing approaches that capture chromatin accessibility, ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) [21] is one of the most popular approaches because of its high sensitivity and the relatively easy and fast lab protocol. ATAC-seq employs a hyperactive form of the *Tn5* transposase that randomly cuts and ligates DNA sequences. These insertion events are more likely to occur in active and accessible chromatin regions, and the ligation activity can be co-opted to ligate sequencing adapters to the fragment, so that a single step results in a sequenceable library, with a sequenced fragment a proxy of chromatin accessibility across the genome. Recently, the development of single-cell ATAC-seq (scATAC-seq) approaches has extended the measurement of chromatin accessibility to the level of the individual cell, revealing sub-population cell-level heterogeneity covered in bulk ATAC-seq assays.

A core step of ATAC-seq and scATAC-seq analysis is the identification of accessible regions (i.e. peak calling) from aligned reads. Several factors contribute to making this process of peak calling complicated. First, the lack of input controls for ATAC-seq (i.e. random cleavage signals by *Tn5* on naked DNA) make it necessary to develop methods not requiring controls, which is challenging as the signal-to-noise ratio varies between experiments. Second, it is difficult to capture low coverage signals, both when analyzing bulk datasets of low sequencing depth, as well as single-cell data with low cell number. Finally, the sparsity of ATAC-seq signals add extra complications for identifying and distinguishing rare signals, especially in single-cell datasets comprising rare cell populations.

In recent years, researchers have developed several novel algorithms or improved existing ones for ATAC-seq peak calling. Traditional methods, such as MACS2, F-seq and Genrich, detect open chromatin regions by either piling up short fragments of size within a defined range, or using a shift-extend approach to count the insertion events smoothed by a defined extension size. Recently, HMMRATAC [53], a HMM-based method is proposed to specifically peak call on ATAC-seq, defining three hidden states based on the fragment size according to integer number of nucleosomes. Other methods are proposed for imputation and denoise for data with low sequencing depth, such as AtacWorks [81] that uses a supervised learning approach to predict denoised signals from noisy input based on training on abundant data. However, none of these methods take into account the two important information, fragment size and coverage into a mechanistic modelling approach, which we believe has the potential to produce peak calls in higher resolution and assist with biological explainability.

To address these challenges, I developed EpiCall, a light-weight statistical model that directly models the mechanics of transposase insertion to make optimal use of the information represented by ATAC-seq sequencing reads to sensitively identify accessible chromatin regions at base-pair resolution. EpiCall explicitly models the probability of observing a fragment as the joint probability of i) two transposase insertion events taking place at both ends of the fragment, and ii) the fragment being sequenced. In the implementation of the model, we employed a hidden

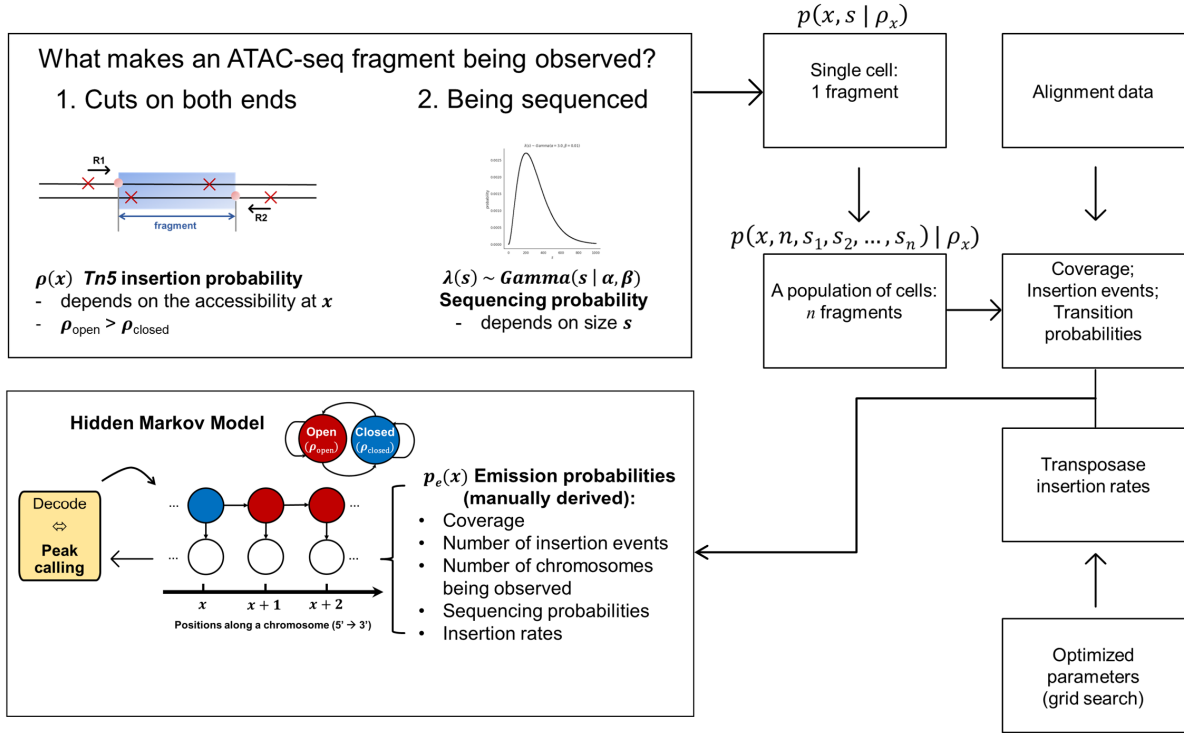


Figure 3.1: EpiCall model structure.

Markov Model with two hidden states, *open* for active chromatin regions and *closed* for inactive heterochromatin regions, which are associated with a high and a low insertion rate of transposase, respectively. We believe the structure of the model can capture enriched regions more sensitively without the need for signal denoise or hard filtering on fragment sizes prior to peak calling.

In this Chapter, we aim at demonstrating the performance of EpiCall in comparison with state-of-art current methods on various bulk and single-cell ATAC-seq datasets to prove its advantage for high-resolution peak calling and detecting cell-type specific genetics information. We also show that EpiCall is robust for producing highly accurate peaks from data of varied sequencing depths when the sequencing quality is low or the number of cells is insufficient.

3.2 EpiCall

3.2.1 The Model Concept

EpiCall takes paired-end ATAC-seq or scATAC-seq sequencing alignment data in BAM or 10X-style fragments (.tsv) formats as input and retains reads which are correctly paired and of high mapping quality. EpiCall is implemented using a hidden Markov Model (HMM) structure in which we define two hidden states, *open* for chromatin regions with high accessibility and a higher transposase insertion rate ρ_{open} , and *closed* for chromatin regions with low accessibility and a lower transposase insertion rate ρ_{closed} . We assume a simplified scenario that each position in the genome is either *open* or *closed* (Figure 3.1). We model the mechanism for the generation of ATAC-seq fragments and denote the HMM state for each genomic position. The aggregated *open* states will be our peak calls.

3.2.2 Derivation of the Probabilistic Model

An ATAC-seq fragment in an ATAC-seq experiment is observed because 1) two neighbouring transposase insertion events occur, one at either end, and 2) the resulting fragment is successfully sequenced. We propose a probabilistic model for the observation of fragments with varying sizes in a population of cells.

First, we focus on the formation of fragments on a single chromosome. At any chromosome position x , a fragment of size s that starts at x is the result of two transposase insertions, one at position x and one at $x + s$. Let $\rho(x)$ be the probability of transposase insertion at x and $\lambda(s)$ be the probability that a fragment of size s ends up being sequenced; then the probability density of observing a size- s fragment starting at x is

$$p(x, s) = \rho(x)[1 - \rho(x + 1)] \cdots [1 - \rho(x + s - 1)]\rho(x + s)\lambda(s) \quad (3.1)$$

$$\approx \rho(x) \exp \left[- \int_{y=x}^{x+s} \rho(y) dy \right] \rho(x + s)\lambda(s) \quad (3.2)$$

where we made a continuous-space approximation to convert the product into an integral. In the simplified case where $\rho(x) = \rho$ (i.e. the probability of transposase insertion is uniform across the genome segment $[x, x + s]$), the probability

(density) becomes

$$p(x, s \mid \rho) = \rho^2 e^{-\rho s} \lambda(s) \quad (3.3)$$

Next, we model $\lambda(s)$. Sequencing machines tend to not be able to sequence fragments that are either too short or too long. The histogram of sequence fragment lengths is approximately normal but strictly positive [82–84], similar to a Gamma distribution. We therefore assume that the probability that a fragment of size s is sequenced is proportional to $\text{Gamma}(s \mid \alpha, \beta)$. We define λ_0 as the probability that a fragment with optimal size ends up being sequenced. This optimal size is the mode of the Gamma distribution, $s_{\text{opt}} = (\alpha - 1)/\beta$; so that $\lambda(s)$ becomes

$$\lambda(s) = \lambda_0 \left(\frac{\beta e}{\alpha - 1} \right)^{\alpha-1} s^{\alpha-1} e^{-\beta s} \quad (3.4)$$

and it can be checked that indeed $\lambda(s_{\text{opt}}) = \lambda_0$. Inserting Equation 3.4 into Equation 3.3 and integrating out s gives the probability of observing a fragment of *any* size starting at x :

$$p(x \mid \rho) = \rho^2 \lambda_0 \left(\frac{\beta}{\beta + \rho} \right)^\alpha \frac{\Gamma(\alpha)}{\beta} \left(\frac{e}{\alpha - 1} \right)^{\alpha-1} \quad (3.5)$$

The conditional probability of an observed fragment size, given that it starts at x and given a constant insertion rate ρ , also follows a Gamma distribution:

$$p(s \mid x, \rho) = \frac{p(x, s \mid \rho)}{p(x \mid \rho)} = \frac{(\beta + \rho)^\alpha}{\Gamma(\alpha)} e^{-(\beta + \rho)s} s^{\alpha-1} = \text{Gamma}(s \mid \alpha, \beta + \rho) \quad (3.6)$$

In other words, according to this model, the process of transposase insertions at rate (or per-base probability) ρ modifies the ‘natural’ fragment size distribution $\text{Gamma}(\alpha, \beta)$, which would result from sequencing a library with a hypothetical uniform fragment length distribution and which has mean fragment length α/β , into a fragment size distribution $\text{Gamma}(\alpha, \beta + \rho)$ with shorter fragments, with mean length $\alpha/(\beta + \rho)$.

Many chromosomes

The next step is to model the observations of fragments from a population of cells. Still we consider fragments starting at a single genomic position x . The N chromosomes all have the same probability $p(x) = p(x|\rho)$ of a fragment starting at x , so that the total count $n(x) = n$ of fragments starting at x obeys a binomial distribution $\text{Binom}(N, p(x))$. Since $p(x) \ll 1$ we can approximate it with a Poisson distribution with rate $Np(x)$. The probability of observing n fragments of sizes $s_1, s_2, \dots, s_n$ starting at x is then

$$\begin{aligned}
 p(n, s_1, \dots, s_n | \rho) &= \frac{(Np(x | \rho))^n}{n!} e^{-Np(x|\rho)} \prod_{i=1}^n p(s_i | x, \rho) \\
 &= \frac{(Np(x | \rho))^n}{n!} e^{-Np(x|\rho)} \prod_{i=1}^n \frac{p(x, s_i | \rho)}{p(x | \rho)} \\
 &\propto e^{-Np(x|\rho)} \prod_{i=1}^n p(x, s_i | \rho) \\
 &= e^{-Np(x|\rho)} \prod_{i=1}^n \rho(x) \rho(x + s_i) \prod_{y=x+1}^{x+s_i-1} [1 - \rho(y)] \quad (3.7)
 \end{aligned}$$

where $p(x, s_i | \rho)$ is the conditional probability of observing a fragment of size s_i starting at x , and we dropped all factors independent of ρ .

We now extend this expression for the fragments observed to start at a single position x , to the configuration of fragments observed on a large segment $[-S, S]$:

$$\begin{aligned}
 p(\text{fragments} | \rho) &= \prod_{x=-S}^S e^{-Np(x|\rho)} \prod_{i=1}^{n(x)} \rho(x) \rho(x + s_{i,x}) \prod_{y=x+1}^{x+s_{i,x}-1} [1 - \rho(y)] \\
 &\approx e^{-N \int_{x=-S}^S p(x|\rho)} \prod_{x=-S}^S \rho(x)^{k(x)} e^{-K(x)\rho(x)} \quad (3.8)
 \end{aligned}$$

where we approximated $\sum p(x) \approx \int p(x)$ and $1 - \rho(x) \approx e^{-\rho(x)}$, and we introduced $k(x)$, the number of fragments that start or end at x , and $K(x)$, the total physical coverage at x .

Single rate change

We then consider a special case with a single rate change: $\tau(x) = \rho_1$ for $x < 0$, and $\tau(x) = \rho_2$ for $x \geq 0$. In this case $p(x | \tau)$ depends in a complicated way

on ρ_1, ρ_2 and x . However, it is clear that

$$\begin{aligned} p(x | \tau) &\approx p(x | \rho_1) && \text{for } x \ll 0; \\ p(x | \tau) &\approx p(x | \rho_2) && \text{for } x \gg 0 \end{aligned} \quad (3.9)$$

For x close to 0 the function $p(x|\tau)$ interpolates between $p(x|\rho_1)$ and $p(x|\rho_2)$. Introduce the function $\eta(x) = \eta_{\rho_1, \rho_2}(x)$ so that

$$p(x|\tau) = [1 - \eta(x)]p(x|\tau(x)) + \eta(x)p(x|\tau'(x)) \quad (3.10)$$

where $\tau'(x) := \tau(-x)$ is the function that is equal to ρ_2 if $\tau(x) = \rho_1$ and vice versa, and with $p(x|\tau(x))$ we denote the probability of a segment starting at x , given that the insertion rate is equal to $\tau(x)$ *across the whole genome*, and similarly for $p(x|\tau'(x))$. Note that $\eta_{\rho_2, \rho_1}(x) = \eta_{\rho_1, \rho_2}(-x)$.

Because of Equation 3.9 we find that $\eta(x) \rightarrow 0$ for $x \ll 0$ and $x \gg 0$. That means that we can write

$$\int_{x=-S}^S p(x | \tau) = Sp(x|\rho_1) + Sp(x|\rho_2) + R(\rho_1, \rho_2) \quad (3.11)$$

where $R(\rho_1, \rho_2)$ is a remainder term that does not depend on x .

Two rates, multiple rate changes

Now assume that $\rho(x)$ is a piecewise constant function that only takes on the two values ρ_1 and ρ_2 , then substituting into Equation 3.8 results in

$$p(\text{fragments}|\rho) = \prod_i e^{NR(x_i)} \prod_{x=-S}^S e^{-Np(x|\rho(x))} \rho(x)^{k(x)} e^{-K(x)\rho(x)} \quad (3.12)$$

where i runs over the switch-points where ρ changes value from ρ_1 to ρ_2 or vice versa, and $R(x_i) = R(\rho_1, \rho_2)$ or $R(\rho_2, \rho_1)$ as appropriate. Because $\eta_{\rho_2, \rho_1}(x) = \eta_{\rho_1, \rho_2}(-x)$ we also find that $R(\rho_1, \rho_2) = -R(\rho_2, \rho_1)$, and since we may assume that the genome is in compacted form and subjected to the same low insertion rate ρ_1 at both ends, the $R(\rho_1, \rho_2)$ terms precisely cancel the $R(\rho_2, \rho_1)$ terms, and we get the final expression

$$p(\text{fragments}|\rho) = \prod_{x=-S}^S e^{-Np(x|\rho(x))} \rho(x)^{k(x)} e^{-K(x)\rho(x)} \quad (3.13)$$

3.2.3 Implementation With hidden Markov Model

We define a simple HMM of two hidden states, with the Markov chain of length equivalent to the size of the chromosome being observed. The time steps of the Markov chain in the original definition of HMM correspond to the linear genomic sequence along the chromosome. We denote the two hidden states as $Q = \{q_1, q_2\}$ corresponding to ρ_1 and ρ_2 , respectively. An HMM implementation of EpiCall should build the probabilistic relationship between i) the path of hidden states encoding the transposase insertion rate $\rho(x)$ along the chromosome, and ii) the observations of the counts and sizes of fragments along the chromosome.

We need to first derive the expression of emission probabilities. Since our primary focus is to infer the optimal path, rather than infer parameters by optimizing the full likelihood, we can drop all factors that do not depend on $\rho(x)$, since the only parameter that changes depending on the choice of path is $\rho(x)$. We have derived the probability of the observations in so far as that depends on $\rho(x)$ in Equation 3.13; the factors that depend on the insertion rate ρ at position x constitutes the emission probability at that position:

$$p_e(x) = e^{-Np(x|\rho(x))} e^{-K(x)\rho(x)} \rho(x)^{k(x)} \quad (3.14)$$

Now we can summarise the full HMM with the following components:

- $Q = \{q_1, q_2\}$ - 2 hidden states, corresponding to *Tn5* insertion rates ρ_1 and ρ_2 , respectively.
- $A = \{a_{ij}\}$ - the transition probability matrix, in which each a_{ij} represents the probability of moving from state i to state j , s.t. $\sum_{j=1}^2 a_{ij} = 1, \forall i$.
- $B = [p_e(x | \rho_1), p_e(x | \rho_2)]$ - the emission probabilities along the path conditioned on ρ_1 and ρ_2 , respectively, in which $p_e(x | \rho_i)$ follows Equation 3.14.
- $\pi = \{\pi_1, \pi_2\}$ - an initial probability distribution over the two states, in which π_i is the probability that the Markov chain starts in state i , s.t. $\sum_{i=1}^2 \pi_i = 1$.

We further name the two hidden states *open* and *closed* corresponding to insertion rates ρ_{open} and ρ_{closed} , which represent the simplified binary classification of each genomic position: accessible and inaccessible. As mentioned above, the emission probabilities encode the likelihood of the observations given a path; the transition probabilities encode the prior distribution over paths (and encode e.g. the density and size distribution of peaks).

3.2.4 Calling Peaks With the Viterbi Algorithm

We use the Viterbi algorithm [85] to find the best path of a single best state annotated for every genomic position. We then merge consecutive *open* states into regions without any filtering to provide the raw peak calls. Therefore, the peaks vary in size.

3.2.5 Defining Peak Score

To assign a score to each peak as a measure of ‘peak quality’, we take the maximum of the log odds ratio across all positions within a peak, using the probabilities from the final best path of the Viterbi algorithm, i.e. for an individual peak, its raw peak score is calculated as:

$$score_{raw}(\text{PEAK}) = \max(\log \frac{p(q_{open} | x)}{p(q_{closed} | x)}), \quad x \in \{\text{all positions in PEAK}\} \quad (3.15)$$

In order to facilitate further peak filtering and comparisons between peak callers, the raw peak score is then rescaled between 0 and 1 across peaks of each chromosome, respectively, in order to obtain the final peak score.

For the evaluation of peak calling performance in Equation 3.3.1, we use varied thresholds of these scores to generate the point-wise precision-recall curves (PRC), receiver operating characteristic curves (ROC) and calculate the area under the PRC (AUPRC) and area under the ROC (AUROC).

3.3 Results

The focus of this section is an extensive evaluation of EpiCall, with the aim to prove its advantage of the peak calling performance on various publicly available bulk and single-cell ATAC-seq datasets.

3.3.1 Accurate and sensitive detection of cell type-specific open chromatin regions

Because HMM computes path probabilities and not true likelihoods, standard (Baum-Welch) training of this model is not appropriate. Instead I performed a large-scale grid search to find the optimised combination of ρ_{open} and ρ_{closed} . I used an ATAC-seq dataset of 50K cells of the GM12878 cell line (GEO: GSM1155957) [21], one of the first bulk ATAC-seq datasets. To ensure sufficient data for hyper-parameter search, I merged three replicates of the GM12878 50k samples (Replicate 2-4; Replicate 1 is discarded as it has much higher read depth than the other replicates). The performance of each hyper-parameter setting was measured by the overlap of base pairs between the peak calls and the reference ‘gold-standard’ sets, constructed by taking the candidate CREs (cCREs) of GM12878 cell line in SCREEN [7] as real positives (active regions), and merging the repressive states from the 18-state ChromHMM annotations on the GM12878 cell line as real negatives (inactive regions) (Methods 3.6.3). For point-wise calculation for AUPRC, AUROC and F1 scores, I set up a range of cut-offs on peak scores to produce different subsets of peaks (Figure 3.2A). Computation details are in Section 3.6.3. Since peaks cover only a small fraction of the genome (≤ 0.1), the AUPRC is a more appropriate statistic for the estimation of hyper-parameters, with its highest value corresponds to $\rho_{open} = 0.0015$ and $\rho_{closed} = 0.0008$ selected as the optimised hyper-parameter settings. Summary statistics shows that number of fragments overlapping with the peak calls under the optimised settings constitutes 23.4% of the total number of fragments (Figure 3.2C). As expected, these peak-overlapping fragments are more enriched in size below 200bp, less enriched in larger size and show less obvious enrichment around integer number of nucleosomal sizes in comparison with non-peak-overlapping fragments. However, there are still non-negligible percentage of peak-overlapping fragments larger than 200bp (Figure 3.2B).

To evaluate hyper-parameter optimisation in terms of the ability to generalize, I applied the selected hyper-parameters and performed the same analysis on four ENCODE cell lines, K562 (an erythro-leukemic cell line), HepG2 (an hepatocellular

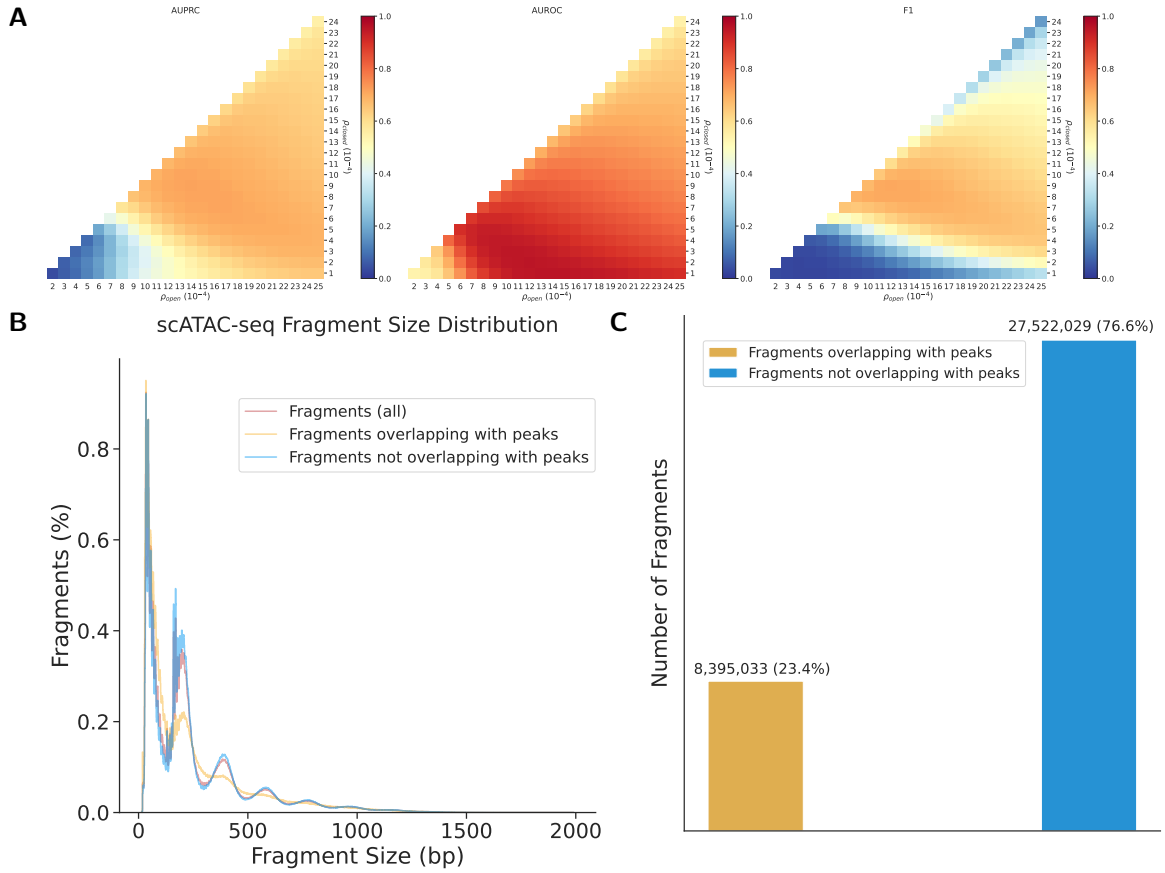


Figure 3.2: Hyper-parameter optimisation for EpiCall using GSE47753 GM12878 50K as training data. **A.** AUPRC, AUROC and F1 scores of EpiCall peak calls on GM12878 training data [21] under different combinations of ρ_{open} and ρ_{closed} in hyper-parameter search space, referencing the GM12878 cell-line-specific SCREEN cCREs [7] as real positives and the merged the repressed states of the 18-state ChromHMM model as real negatives; **B.** Fragment size distribution of GM12878 training data, colored by all fragments (red), fragments overlapping with peaks (yellow) and fragments not overlapping with peaks (blue), in which the peaks calls are produced by EpiCall under the selected hyper-parameters from **A**; **C.** Barplots showing the number of fragments overlapping and not overlapping with the peak calls produced by EpiCall as in **B**.

carcinoma cell line), IMR-90 (a normal human foetal lung fibroblast cell line) and Panc1 (a pancreatic carcinoma cell line) (see Appendix 1 for the source of data) along with the previous GM12878 (a B-lymphocyte cell line) dataset used for hyper-parameter tuning. I compared the performance of EpiCall against three existing peak callers: MACS2 [49], the current standard peak caller for ChIP-seq and ATAC-seq; HMMRATAC [53], an HMM-based ATAC-seq specific peak caller; and LanceOtron [55], a deep-learning peak caller for various NGS assays including ChIP-seq and ATAC-seq.

In the summary statistics of peak calls across the five cell lines (Table 3.1), EpiCall tends to produce peaks in smaller size while HMMRATAC and LanceOtron consistently produce large peaks. For the evaluation of peak calls (Figure 3.3), in addition to the references used in the hyper-parameter optimisation stage (Figure 3.2), I included a second reference set defining the real positives as the active states of 18-state ChromHMM annotations (Methods 3.6.3) and the real negatives still as the ChromHMM inactive states. Both reference sets are cell-line specific. The reason for using the two reference sets is to evaluate the precision and recall under different inclusion of real positives, in which the SCREEN cCREs are more stringent for open-chromatin references with support from DNase-hypersensitive sites and CTCF in addition to the broad histone modifications that ChromHMM were modeled on. Across all five cell lines (i.e. the GM12878 dataset for hyperparameter search and the four ENCODE cell lines), EpiCall generally achieves high AUPRC and consistently reaches higher precision under a wide range of recall values in reference with the SCREEN cCREs (active) - ChromHMM (inactive) annotations, while performing less optimal to HMMRATAC in K562 and achieving lower recall in HepG2 in reference with the ChromHMM (active) - ChromHMM (inactive) annotations in K562 and HepG2. These results suggest the hyper-parameters selected from grid search are generally applicable to other datasets in terms of sensitive identification of cell line-specific active regulatory regions while avoiding inactive regions. Despite of the encouraging performances, it is again worthwhile to note that these references, especially the real positives, are only broad estimations from functional genomic assays of varying quality, which can be too rough for evaluating ATAC-seq peak calls in terms of high-resolution requirements ¹. In the next sections, I present evaluations measurements for the correspondence of the positions between the actual signal enrichment of functional genomics and the ATAC-seq peak calls.

¹The ‘broad estimations’ here refers to reference sets modeled from peak calls of ChIP-seq for histone modifications. These peak calls are produced by MACS2 broad peak call settings suitable for histone binding patterns, which cover wider regions and are at lower resolution than ATAC-seq data digested by transposases. This also explains the observation in Figure 3.3 that when calculating confusion matrices in terms of overlapping base pairs, the peaks across all peak callers in the majority of the assessed cell lines only reaches to a maximum recall below 0.8.

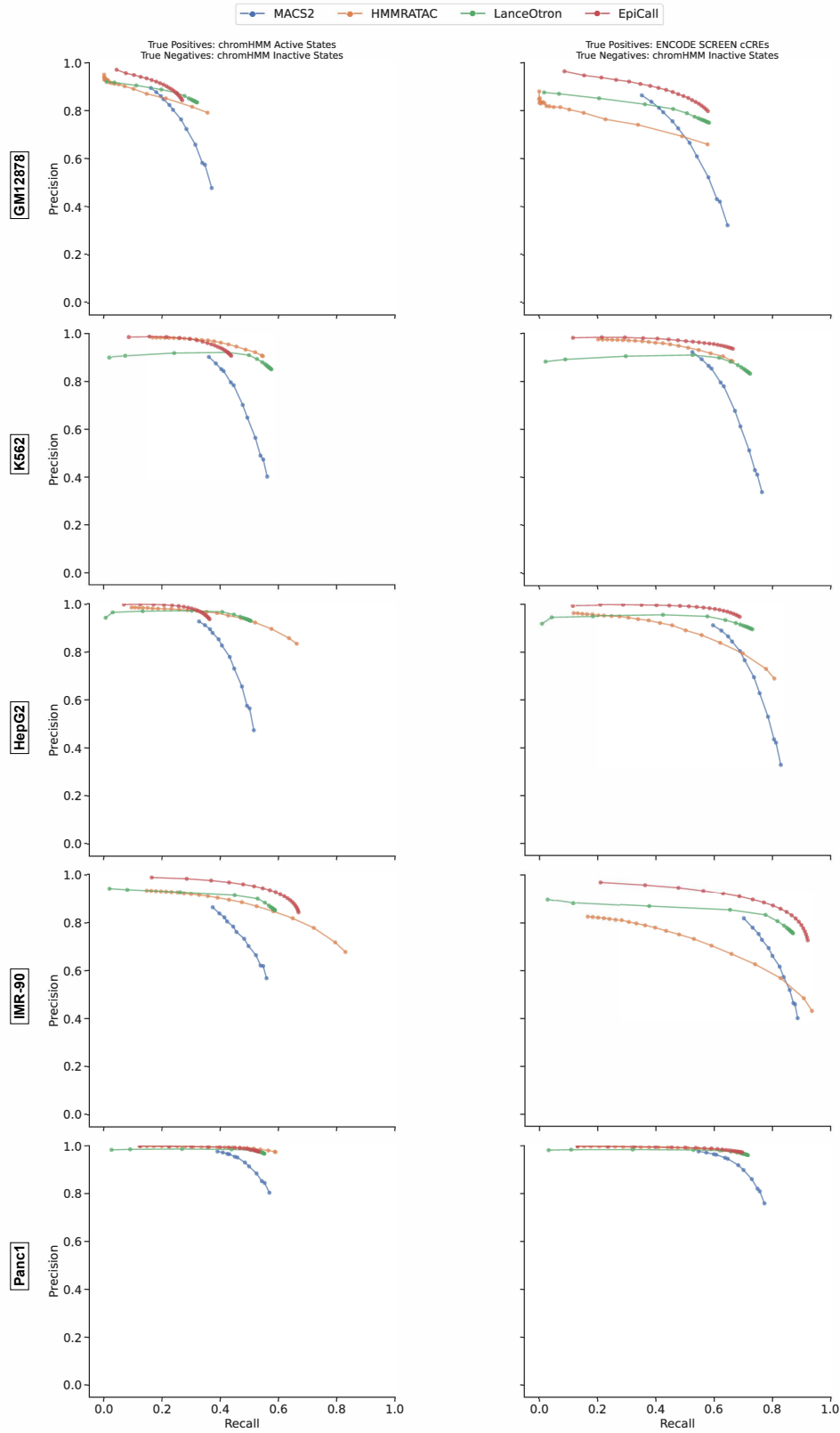


Figure 3.3: Precision-Recall curves (PRC) of peak calls produced by MACS2 (default setting, $q=0.05$), HMMRATAC (unfiltered peaks), LanceOtron (overall peak score cut-off at 0.5) and EpiCall (unfiltered peaks) across five bulk ATAC-seq datasets, evaluated using two sets of 'ground truth' references: **left** - real positives as the active states of 18-state ChromHMM annotations and real negatives as the inactive states of 18-state ChromHMM annotations; **right** - real positives as the ENCODE SCREEN cCREs [7] and real negatives as the inactive states of 18-state ChromHMM annotations excluding the ENCODE SCREEN cCREs. Both references are cell-line specific and the calculations are performed on the overlapping base pairs.

Cell line	Peak Caller	No.Peaks	Mean Peak Size	Median Peak Size	Fraction of Genome	No.Unique Peaks
GM12878	MACS2	51,035	761	660	0.0126	6,729
	HMMRATAC	33,596	1,507	1311	0.0164	0
	LanceOtron	41,353	1030	937	0.0138	88
	EpiCall	62,405	563	360	0.0114	11,117
K562	MACS2	84,972	760	649	0.0209	30,745
	HMMRATAC	33,324	1,552	1241	0.0167	32
	LanceOtron	64,991	1047	951	0.0201	2,766
	EpiCall	65,573	619	340	0.0131	3,433
HepG2	MACS2	90,526	718	605	0.0210	29,492
	HMMRATAC	58,189	1,657	1,361	0.0312	844
	LanceOtron	54,117	1066	973	0.0187	236
	EpiCall	74,833	524	300	0.0127	3,626
IMR-90	MACS2	141,238	698	610	0.0319	19,950
	HMMRATAC	105,278	1,958	1,611	0.0667	57
	LanceOtron	106,247	907	789	0.0312	274
	EpiCall	132,540	909	600	0.0390	9,664
Panc1	MACS2	85,402	758	630	0.0210	3,1810
	HMMRATAC	37,048	1,766	1,421	0.0212	47
	LanceOtron	59,710	1,078	1,008	0.0209	3,378
	EpiCall	69,416	664	350	0.0149	3,352

Table 3.1: Summary statistics of peak calls across five cell lines. Peak calling methods are set as default: MACS2: $-q \ 0.05$; HMMRATAC: $\text{score} \geq 0$; LanceOtron: overall peak score ≥ 0.5 ; EpiCall: unfiltered peaks.

3.3.2 EpiCall sensitively recovers TF binding and functional histone modification signals

Functional genomic annotations such as ChromHMM, SegWay, hiHMM and ENCODE SCREEN cCREs provide broad and low-resolution chromatin marks that can be noisy and inaccurate, due to the fact that these models take primarily histone modifications into consideration (except for 25-state chromImpute which considers imputed DNase-seq and SCREEN which includes CTCF ChIP-seq). The broad peak calls from histone ChIP-seq and the lack of TF ChIP-seq data potentially limit the ability of the annotations from accurately capturing the enhancer regions, whose binding sites are sequence-specific in the genome. As a result, the identification of TF binding signals is an important criterion for evaluating ATAC-seq peak calls in a cell-line specific and high-resolution way. Theoretically, good ATAC-seq peak calls should sensitively detect open chromatin regions binding by transcriptional regulatory elements such as CTCF and cell-type specific TFs. The performance can be reflected by (1) the recovery of the exact transcription factor binding sites (TFBS) by ATAC-seq peak calls; and (2) the overlap between ATAC-seq peak calls

and TF ChIP-seq peaks on the same sample.

In this section, I will demonstrate the evaluations on the two criteria above.

Motif enrichment analysis of representative TFs

To systematically investigate the capability of EpiCall to accurately capture important motifs of open-chromatin and enhancers, I performed motif enrichment analysis on ATAC-seq peak calls using HOMER [86] (Methods 3.2) on peak calls produced by EpiCall, MACS2, HMMRATAC and LanceOtron on three out of the five ATAC-seq cell lines evaluated in Section 3.3.1 and selected a subset of representative motifs [87] for demonstration: CTCF is selected across cell lines as a signature TF of active transcription. For cell-line specific TFs, I selected RUNX2 and IRF8 for GM12878, GATA2 and NFE2 for K562, and HNF4A and FOXA3 for HepG2. SP1 is used as a neutral motif that is expected to see no enrichment across cell lines. I used the position weight matrices (PWMs) of the entire JASPAR2022 [60] as the known motif database to search for motif enrichment, and in Table 3.2 report the significance of enrichment (log P -value and the enrichment itself (fold enrichment, % of Targeted peaks with motif presence, % of background sequences with the motif). The results demonstrate that peak calls made by EpiCall tend to have the highest fold enrichment and lowest P -value in CTCF and most of the cell-line specific motifs across GM12878, K562 and HepG2, except that in HepG2 EpiCall produce marginally less significant P -value than MACS2 on the enrichment of HNF4A and FOXA3. HMMRATAC and LanceOtron consistently achieved lower fold enrichment and less significance across cell lines and motifs.

Sensitive positioning of peak calls map with TF and Histone ChIP-seq signal enrichment

In addition to quantitative motif enrichment analysis scanning the covered region of all peaks for the PWMs of motifs, it is equally important to investigate how sensitive the position of peak calls are in line with the signal enrichment of TF binding and functional histone modifications. In contrast to the precise prediction of binding sites from ChIP-seq, MACS2 struggles to find TF binding sites such

Cell line	JASPAR Motif	Peak Caller	log P -value	% of Targets	% of Backgrounds	Fold Enrichment
GM12878	CTCF (MA0139.1)	MACS2	-3679.00	27.26	11.96	2.28
		HMMRATAC	-1800.00	35.07	18.15	1.93
		LanceOtron	-2543.00	30.32	15.38	1.97
		EpiCall	-8538.00	30.53	9.71	3.14
	RUNX2 (MA0511.2)	MACS2	-1235.00	19.75	11.45	1.72
		HMMRATAC	-447.20	28.12	19.80	1.42
		LanceOtron	-898.80	22.81	14.44	1.58
		EpiCall	-1470.00	16.63	9.03	1.84
	IRF8 (MA0652.1)	MACS2	-871.30	2.84	0.65	4.37
		HMMRATAC	-362.80	3.627	1.19	3.04
		LanceOtron	-637.7	3.06	0.84	3.64
		EpiCall	-1089.00	2.47	0.47	5.26
	SP1 (MA0079.5)	MACS2	-130.30	34.54	30.95	1.12
		HMMRATAC	-109.30	42.49	37.73	1.13
		LanceOtron	-156.50	40.84	36.37	1.12
		EpiCall	-201.50	29.04	25.15	1.15
K562	CTCF (MA0139.1)	MACS2	-5543.00	29.49	13.92	2.12
		HMMRATAC	-1082.00	39.93	27.50	1.45
		LanceOtron	-3221.00	33.97	18.96	1.79
		EpiCall	-6926.00	34.36	14.24	2.41
	GATA2 (MA0036.3)	MACS2	-1531.00	23.59	15.49	1.52
		HMMRATAC	-636.60	32.98	23.89	1.38
		LanceOtron	-1108.00	27.54	18.99	1.45
		EpiCall	-1621.00	17.98	9.96	1.80
	NFE2 (MA0841.1)	MACS2	-816.20	3.76	1.53	2.46
		HMMRATAC	-334.30	5.31	2.62	2.03
		LanceOtron	-593.7e	4.34	1.92	2.30
		EpiCall	-837.20	3.48	1.16	3.00
	SP1 (MA0079.5)	MACS2	-89.43	36.54	34.16	1.06
		HMMRATAC	-36.59	60.86	58.53	1.04
		LanceOtron	-90.64	46.70	43.80	1.07
		EpiCall	-15.41	38.27	37.22	1.03
HepG2	CTCF (MA0139.1)	MACS2	-5647.00	28.18	13.42	2.10
		HMMRATAC	-2,184.00	37.96	24.84	1.53
		LanceOtron	-2522.00	33.27	19.57	1.70
		EpiCall	-7122.00	29.54	11.72	2.52
	HNF4A (MA1494.1)	MACS2	-1166.00	4.16	1.54	2.70
		HMMRATAC	-592.50	6.32	3.31	1.91
		LanceOtron	-524.80	25.26	19.27	1.31
		EpiCall	-999.90	3.35	1.06	3.16
	FOXA3 (MA1683.1)	MACS2	-1070.00	19.84	13.73	1.45
		HMMRATAC	-380.30	33.50	28.00	1.20
		LanceOtron	-742.70	24.93	17.93	1.39
		EpiCall	-990.60	11.82	6.89	1.72
	SP1 (MA0079.5)	MACS2	-118.2	35.11	32.51	1.08
		HMMRATAC	-44.98	51.40	49.39	1.04
		LanceOtron	-111.2	48.17	44.84	1.07
		EpiCall	-58.84	33.44	31.47	1.06

Table 3.2: Motif enrichment statistics of ATAC-seq peak calls in cell lines GM12878, K562 and HepG2. Position Weight Matrices (PWMs) of Representative motifs are chosen from JASPAR2022 [60] databases, including CTCF, motifs specifically enriched in each cell line (differing between cell lines) and a neutral motif (SP1) as control reference. The table reports the significance (log P -value, % of peaks (targets) with motif PWM, % of genomic background sequences with motif PWM, and fold enrichment (% of Targets / % of Backgrounds) of each motif. Enrichment analysis is performed using HOMER [86] (Methods 3.6.4).

as CTCF in high resolution from bulk ATAC-seq data due to the low-resolution signals [88], though it performs better on scATAC-seq data which comes with higher signal-to-noise ratio. Similar challenges are also observed for other TFs as well as promoter/enhancer associated Histone modifications. Peak callers such as MACS2 and HMMRATAC provide a ‘summit call’ option in addition to peak calls, where a ‘summit’ of a peak refers to the position of the highest signal coverage. For these peak callers, the summit serves as a more proper choice than the peak center for predicting the binding sites of TFs and Histones. Although EpiCall has not been provided with a summit call option, our assessments suggest the center of EpiCall peak calls serve as accurate estimations for the binding sites of functional elements (Figure 3.4). To demonstrate the positional accuracy of the peaks, I took the centers of ATAC-seq peak calls by EpiCall, MACS2, HMMRATAC and LanceOtron as well as the summits of MACS2 and HMMRATAC as anchor positions, respectively, and computed the pile-up average profile of ChIP-seq signals surrounding the 500bp window to both sides of the anchor positions. For ChIP-seq data, I took the fold-change over control BigWig signal profiles from ENCODE public database including CTCF, selected cell-line-specific enhancer-binding TFs, and histone modifications H3K27ac, H3K4me3 and H3K4me1, all from the same cell line as the corresponding ATAC-seq (see Appendix 1 for data source). Results on GM12878, K562 and HepG2 (Figure 3.4) suggest the center of ATAC-seq peaks made by EpiCall, without summit calls, sensitively align with the enrichment of CTCF and cell-line specific TF profile and display the typical foot-printing feature of Histone markers profile. Among the six positioning strategies in evaluation (four types of peak centers and two types of summits), EpiCall was only outperformed by the summit calls of HMMRATAC, suggesting EpiCall peak calling alone is capable of sensitively predicting functional binding sites of functional elements.

To have a closer look at the peak calls with regard to their functional identities, I focus on the GM12878 50k peak calls, annotating each peak with its nearest SCREEN cCRE. Peaks with a distance to the nearest cCRE larger than 1000bp are discarded. This leads to the remaining peaks being annotated as one of

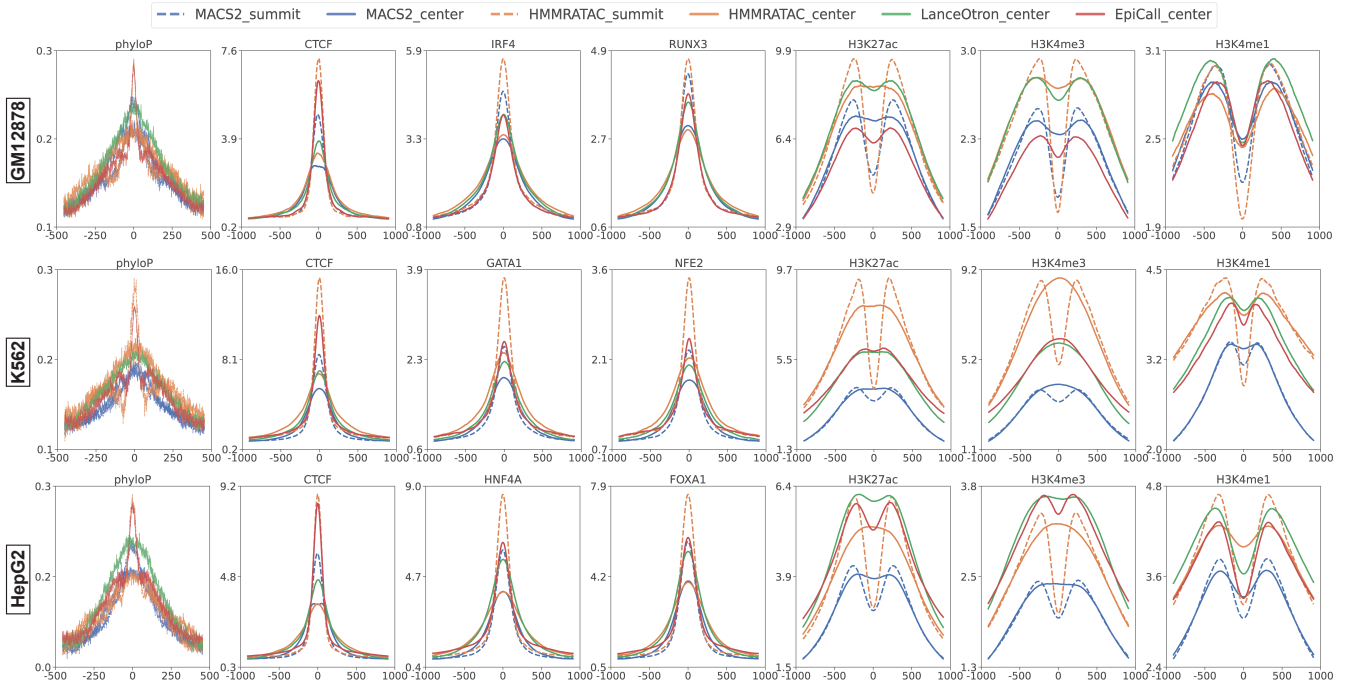


Figure 3.4: Pileup average signal of evolutionary conservation and epigenomics surrounding peak centers or summits, showing phyloP score (column 1) and ChIP-seq of CTCF, representative TFs and histone marks (columns 2-7) surrounding peak centers (solid lines) or summits (dashed lines) made by MACS2 (blue), HMMRATAC (orange) and EpiCall (green) on bulk ATAC-seq data of GM12878 (top), K562 (middle) and HepG2 (bottom).

‘CTCF-only’, ‘PLS/CTCF’, ‘DNase-only’, ‘pELS/CTCF’, ‘dELS/CTCF’, ‘PLS’, ‘pELS’, ‘dELS’². Same as in Figure 3.4, I generated the pile-up signal profile of the functional epigenomics surrounding the peak centers or summits, but instead of averaging the profile across peaks, I displayed the pile-up profile of each peak in a heatmap to show individual signal enrichment (Figure 3.5). Extracting the peaks annotated with ‘CTCF-only’, ‘PLS/CTCF’, ‘pELS/CTCF’ and ‘dELS/CTCF’, the center of the individual peak calls of EpiCall shows a higher enrichment of the CTCF ChIP-seq signals comparable to the summit calls of HMMRATAC and surpassing all other peak calls. Demonstrations of other ChIP-seq signals (cell-type specific TFs and Histone modifications) grouped by active SCREEN cCREs are shown in Appendix Figures 2,3,4.

To further investigate the functional characteristics of EpiCall peak calls, I extracted the exclusive peak calls made by EpiCall, MACS2, HMMRATAC and

²PLS: promoter-like signatures; pELS: proximal enhancer-like signatures; dELS: distant enhancer-like signatures

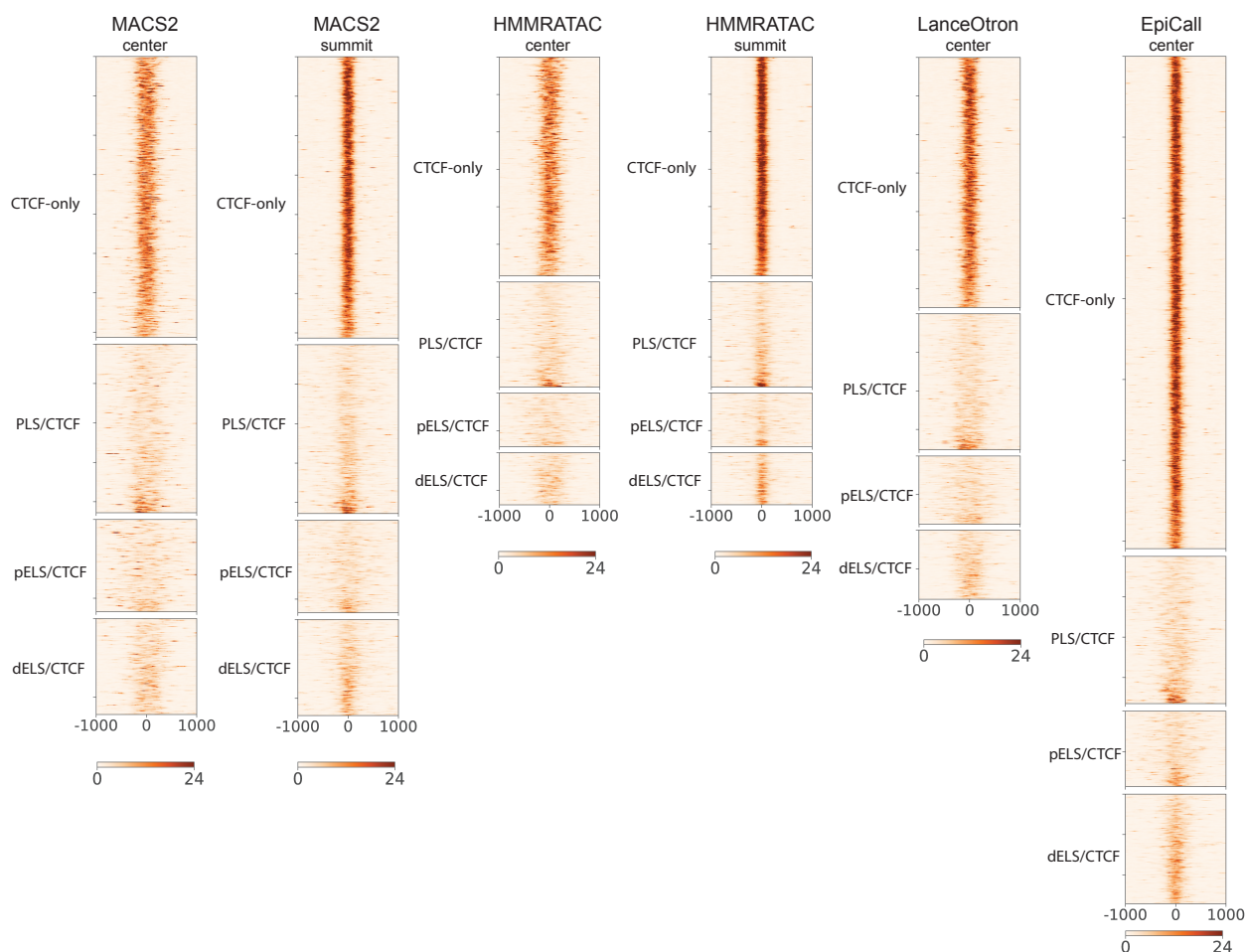


Figure 3.5: EpiCall produces bulk ATAC-seq peak calls sensitively correlated to functional epigenomics. Left to right: Pileup heatmap showing the signals of CTCTF ChIP-seq surrounding MACS2 peak centers, MACS2 peak summits, HMMRATAC peak centers, HMMRATAC peak summits, LanceOtron peak centers, and EpiCall peak centers, all in GM12878 50K dataset. Each column is grouped by ENCODE SCREEN cCREs annotations which are CTCTF-related categories, separated in each column, from top to bottom: CTCTF-only, PLS/CTCF, pELS/CTCF and dELS/CTCF.

LanceOtron, respectively, which have no overlap with any other peak caller. Selecting the highest-scored peaks produced by each peak caller, respectively, I used a powerful interactive visualisation tool for genomics, Multi-Locus-Viewer (MLV) [76], to visualise the epigenomic signals surrounding the center of each individual peak (Figure 3.6). Focusing on chromosome 1, the number of exclusive peaks produced by EpiCall, MACS2, HMMRATAC and LanceOtron is 1024, 583, 0 and 5. From these peaks, I selected the top 10 scored EpiCall peaks (Figure 3.6A), top 10 scored MACS2 peaks (Figure 3.6B), and all 5 LanceOtron peaks (Figure 3.6C) to display

in high resolution. We can see the top exclusive peaks made by EpiCall not only correspond to sharp and narrow enrichment in ATAC-seq signals (red), but also align with the high enrichment of CTCF (orange) (peak 2-6, 8-10) or enhancer TFs such as IRF4 (purple) (peak 6). In contrast, the ATAC-seq signals corresponding to the top MACS2 peaks are ‘flatter’ and less enriched, though peak 2, 5, 8-10 match with the enriched TF signals. Four out of the five LanceOtron peaks have no correspondence to clear ATAC-seq signal enrichment neither within or surrounding the peaks, except for peak 2. None of the LanceOtron peaks have support from observable signal enrichment in CTCF, enhancers or Histone signatures. Notice that the signals of ChIP-seq of Histone modifications, which are in ‘broader’ shape than ATAC-seq and TF ChIP-seq, are likely to be in lower resolution and sensitivity in comparison with ATAC-seq signals, thus become less informative for assessing ATAC-seq peak calls.

3.3.3 High Evolutionary Conservation in EpiCall Peaks

Another characteristic of successful ATAC-seq peak calls is evolutionary conservation. Functional *cis*-regulatory elements in genome are generally highly evolutionarily conserved, which coincides with TF ChIP-seq and ATAC-seq enrichment. The evolutionary conservation across the human genome is quantitatively measured and reported in metrics such as phyloP and phastCons [89–94], both providing base-pair resolution scores representing evolutionary conservation. To benchmark, I took a similar approach to Section 3.3.2, using phyloP score as the query assay and calculate the pile-up average profile surroundings the peak centers (Figure 3.4, 1st column). In all three cell lines, the peak centers called by EpiCall show a sharp enrichment of phyloP score in comparison to MACS2, MMRATAC and LanceOtron, while decreasing to similar low levels from around 50bp away from the center. This provides further evidence the locations of the EpiCall peaks are more evolutionarily constrained and more likely to represent functional regulatory elements.

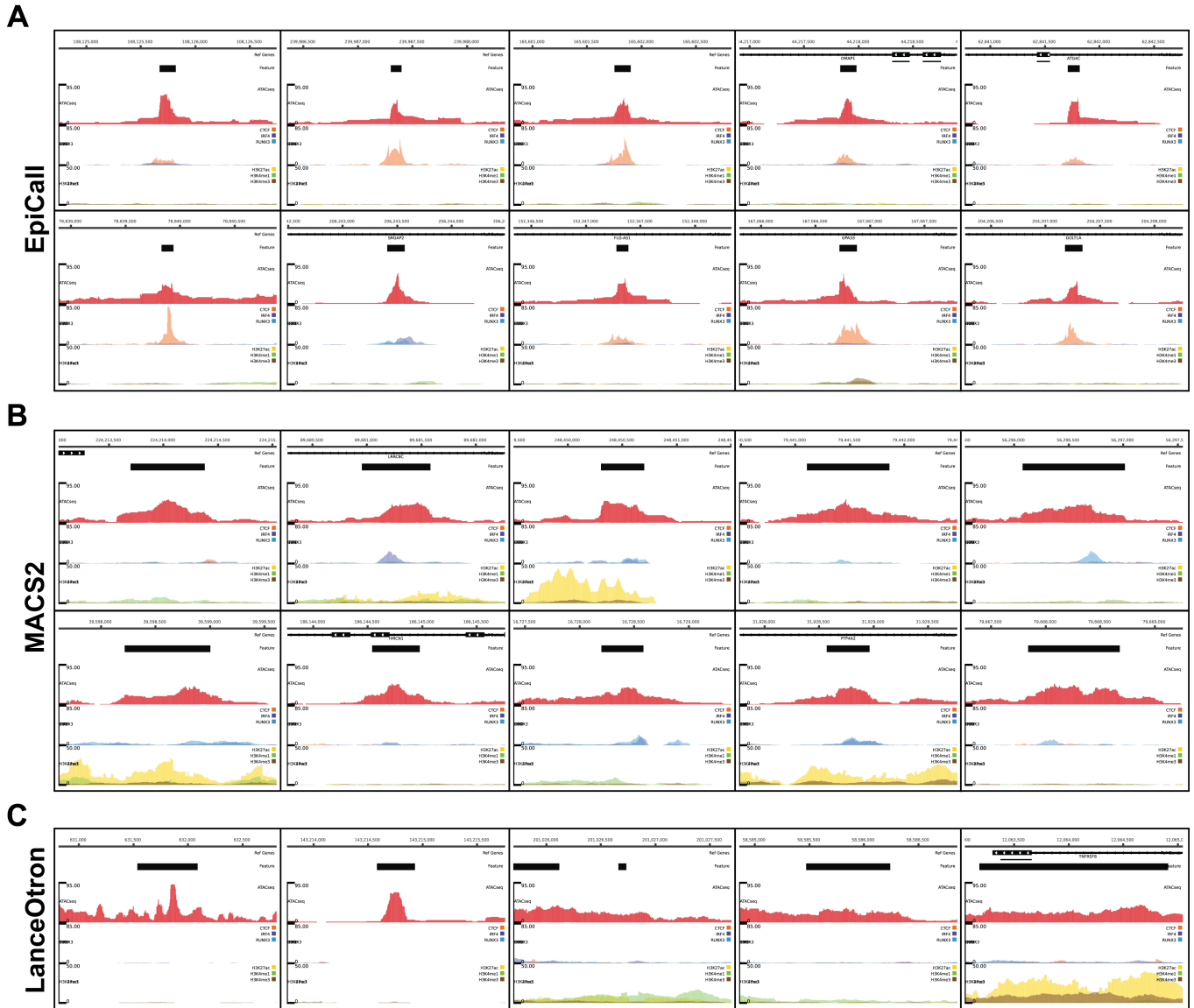


Figure 3.6: Epigenomics signals surrounding top-scored exclusive peak calls produced by EpiCall (A), MACS2 (B) and LanceOtron (C) on GM12878 in chromosome 1. In each thumbnail block, from the top to bottom row, displays the RefSeq gene locus (black arrow); the individual ATAC-seq peak call (black); the ATAC-seq signals used for peak calling (red); the ChIP-seq signals (fold-change over control) of TFs CTCF (orange), IRF4 (purple) and RUNX3 (blue); the ChIP-seq signals (fold-change over control) of Histone modifications H3K27ac (yellow), H3K4me1 (green) and H3K4me3 (brown). Across all thumbnail blocks, The zoom-in scale on the horizontal axis of the genome view are kept the same; the signal range of ATAC-seq, TF ChIP-seq and Histone ChIP-seq are set as (0, 95), (0, 85) and (0, 50), respectively. ^a

^aThe thumbnail plots are arranged by Multiple Locus View (MLV) [76] and available at https://mlv.molbiol.ox.ac.uk/projects/multi_locus_view/7182 upon request.

3.3.4 Robustness to low sequencing depth

The data used in the analyses so far are of relatively high quality with sufficient read depth and cell number, which is however not always achievable in every experiment. In practice, issues including technical operations and rare cell type due to low abundance or difficulty in tissue/sample dissection result into data insufficiency. The varying signal-to-noise ratio and sequencing depth in experimental data bring challenges to peak calling, and thus it has always been a high demand for developing analysis tools robust to varying quality.

To evaluate the performance of EpiCall at low sequencing depth, I generated randomly down-sampled alignment data from the original GM12878 bulk ATAC-seq dataset to challenge the peak callers. From the total 88,228,974 reads, I used `samtools view -s` to randomly sample the reads by 50%, 10%, 5% and 1%, repeating three times with different seeds for each rate. Peak calling was applied using EpiCall with the optimised parameters, MACS2 with three different false discovery rates ($q=0.05$ the default, as well as 0.005 and 0.5), HMMRATAC with the default settings and LanceOtron with overall peak score cut-off at 0.5. For performance evaluation, I reported the Jaccard similarity index ³, recall, precision and F1 scores of the peak calls using two different reference sets. The first reference evaluates the robustness of peak callers by constructing the real positive regions (i.e. accessible) as the peak calls in respective to each peak calling setting on the raw data without down-sampling, and the real negative regions as the complement of the real positives. The second reference is the same as the benchmark reference in Section 3.3.1, using SCREEN cCREs as the real positives and taking the 18-state ChromHMM negative states as the real negatives. Figure 3.7 demonstrates the results, showing the mean and standard deviations of each evaluation statistic at different down-sampling rates, in which EpiCall, MACS2 ($q=0.05$) and LanceOtron were capable to peak call from 50% (4.41×10^7 reads) to as low as 5% (4.41×10^6 reads) of the original data, with EpiCall consistently exceeded the other peak callers

³The aim of using Jaccard similarity index is to penalise false positives, credited to a bioinformatics contest <https://www.synapse.org/#!/Synapse:syn26720920/wiki/615338>

in terms of Jaccard index, recall and F1 score. Noticeably, the superior performance of EpiCall is more significant as the number of reads decrease. HMMRATAC was not able to produce valid peaks ⁴ when the down-sampling rate is as low as 10%. All peak callers failed to produce adequate peak calls at 1% (8.80×10^5 reads) down-sampling rate, but the threshold of read number for valid peak calls has not been fully explored. Overall, the evaluation on down-sampled bulk ATAC-seq indicate EpiCall is by comparison more robust to data with low sequencing depth.

The GM12878 dataset is used for hyperparameter grid search in Section (3.3.1). The evaluation analysis across various ENCODE cell lines prove the optimised hyperparameters are not overfitting in other datasets (Section 3.3.1 and 3.3.2), therefore the conclusions drawn from the down-sampling analysis using the GM12878 dataset is representative. However, it will be more convincing and unbiased to apply the same down-sampling procedure and analysis in other datasets, especially for assessing the generalisability of the results.

3.3.5 Integrating fragment size and coverage benefits peak calling

The model architecture of EpiCall enables integrating both the size and coverage information of ATAC-seq fragments for peak calling. Figure 3.8 shows a qualitative example of the K562 cell line around the gene loci of hemoglobin beta and delta sub-units. I perform peak calling on the entire ATAC-seq data using MACS2, HMMRATAC, LanceOtron and EpiCall. For comparison, I additionally perform peak calling on filtered ATAC-seq data of fragments <150bp, similar to the strategy adopted in Chapter 2. In this example, the region highlighted by the rightmost dashed box has an enrichment of ATAC-seq coverage signals in the entire data and a more significant enrichment in the <150bp filtered data, while being a local minimum of the fragment size within the regions around the peak call. The corresponding sharp enrichment in TAL1 binding signals provides evidence for the peak being a true accessible enhancer, suggesting using the additional information

⁴Peaks labeled with `gappedPeak` in the HMMRATAC output are valid peak calls.

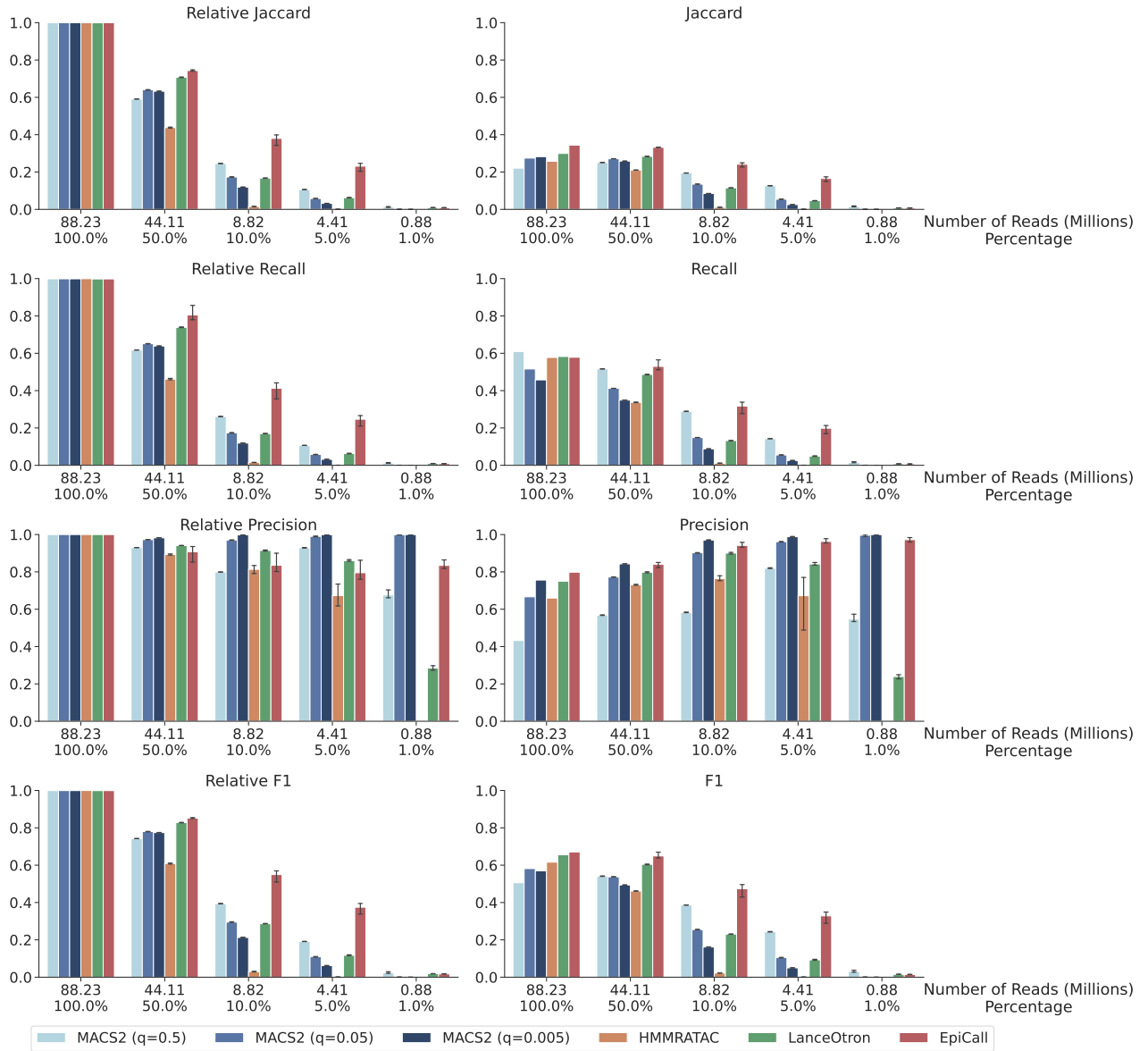


Figure 3.7: Peak calling performance on read down-sampled GM12878 bulk ATAC-seq evaluated by ‘self comparison’ (left) and ‘general performance’ (right). The x -axis labels the number of reads and the corresponding percentage of down-sampling. Notice that HMMRATAC fails to produce valid peak calling at down-sampling percentage $\leq 10\%$. Colors represent the four peak callers, in which MACS2 is performed at three different false discovery rates (q values) 0.005, 0.05 (default) and 0.5. Down-sampling was performed on the original aligned reads by randomly selecting 50%, 10%, 5%, 1% of the reads, repeating for three times at different seeds. For the reference regions used for evaluation, the left plot serves as a ‘self comparison’ by setting the real positive regions as the peak calls in respective to each peak calling setting on the raw data without down-sampling, and the real negative regions as the complement of the real positives. The statistics in self comparison shows the relative Jaccard similarity index, recall, precision and F1 score respective to the raw peak calls. The right plot, complementarily, evaluates the ‘general performance’ of peak calls the same using the cell-type specific ENCODE SCREEN cCREs as real positives and the 18-state ChromHMM negative states as real negatives.

of fragment size distribution assists EpiCall to identify the region as an individual peak from the entire data, while MACS2 is only able to distinguish it from its left neighbouring peak when calling from the <150bp filtered data. The middle highlighted peaks produced exclusively by MACS2 from the entire data is likely to be a false positive, as it has no correspondence to the enrichment in any TF binding despite of low amplifications in H3K4me1 and H3K4me3 signals; In terms of ATAC-seq signals, there is no observable decrease in ATAC-seq fragment size despite of the enrichment in coverage. The leftmost highlighted peak, produced exclusively by MACS2 from the filtered data, is also likely to be a false positive, as no clear enrichment in TF binding provides evidence. However, the region of this peak does has enriched signals of fragments <150bp and a decrease in the mean fragment size, which indicates the mechanistic modeling approach by EpiCall has a superior performance than a simple cut-off size-filtering approach. The conclusion is further supported by the 18-state ChromHMM model which annotates the region of the rightmost peak a single ‘Flanking TSS Downstream’ state and the other two as ‘Weak Enhancer’, suggesting EpiCall can accurately and sensitively identify true peaks with high resolution from noisy input while avoiding false positives.

The comprehensive evaluations thoroughly demonstrate the consistently superior performance of EpiCall over MACS2, HMMRATAC and LanceOtron across several aspects. Specifically, EpiCall sensitively finds regions associated with functional genomic signals such as CTCF and enhancers, as well as shows the potential of producing more robust peak calls on low quality data. The challenge of low quality data is analogous to the problem of peak calling from clusters of insufficient cell number in scATAC-seq. In both situations the challenge is to extract biological information as robustly as possible from sparse data. Chapter 4 follows up on this observation to evaluate if EpiCall advances in information extraction in scATAC-seq peak calling for data of low cell number.



Figure 3.8: K562 ATAC-seq peak calls in reference with functional genomics assays surrounding the hemoglobin gene locus HBB and HBG subunits, showing potential true and false positive peak calls. From top to bottom rows: track 1: ATAC-seq signal coverage; track 2: ATAC-seq signal coverage of fragment size <150bp; track 3: the mean (red line) and standard deviation (red shaded areas) of per base-pair fragment sizes; tracks 4: ATAC-seq peak calls for the entire data by MACS2; track 5: ATAC-seq peak calls for data of fragment size <150bp by MACS2; tracks 6-8: ATAC-seq peak calls for the entire data by HMMRATAC, LanceOtron and EpiCall, respectively; tracks 9-11 (purple): ChIP-seq signals of CTCF, TAL1 and NFE2; track 12-14 (brown): ChIP-seq signals histone marks H3K27ac, H3K4me1 and H3K4me3; track 15: 18-state ChromHMM annotations of K562 (see Appendix 2 for the color reference); track 16: the UCSC RefSeq genes. The two left dashed boxes indicate potential false-positive peaks made by MACS2. The right gray box dashed box highlights a potential high-resolution true-positive peak made by EpiCall from the entire data and MACS2 from the < 150bp data, which is failed to be distinguished in MACS2 and LanceOtron while being missed by HMMRATAC called from the entire data.

3.4 Discussion

3.4.1 High-resolution peak calls enable accurate signal prioritisation of epigenomic profiling

Recognising signals from data of varied levels of quality has long been a challenge in next generation sequencing data analysis. Supervised learning approaches have been trying to infer the adjusted signals in noisy data by training to match high-quality profiles [81], but it is necessary to optimise the ATAC-seq peak calling algorithm itself as an improvement on signal processing. The results in this chapter showed the superior performance of EpiCall to call peaks more robustly on low-quality data utilizing the fragment size distribution in addition to coverage to improve the accuracy and sensitivity when signal is low. The approach for modeling the mechanism of ATAC-seq fragment generation also bypasses the limitation of count-based or shape-based models.

3.4.2 Future work

Exploration of more hidden states and extension to other NGS assays

We believe EpiCall is able to extend to peak call for other NGS types (i.e. DNase-seq, MNase-seq) which produce similar fragment size distributions. In addition, extending the current model to an infinite hidden Markov Model (iHMM) [66], the Bayesian non-parametric version of HMM which automatically selects the optimal number of hidden states, may reveal the multi-facet epigenomics information hidden in ATAC-seq, including but not limited to peak calling, TF footprinting, nucleosome positioning and signal characterisation within gene bodies for the integration with RNA-seq.

Accounting for *Tn5* binding bias

The enzymatic cleavage by *Tn5* on DNA sequences, rather than being purely dependent on the level of chromatin accessibility, has intrinsic sequence bias associated with GC content, which is known to be a major challenge in ATAC-seq footprinting analysis as the bias interferes statistical models in accurately finding

the exact binding sites of the effector proteins. Various footprinting methods have made effort to optimise the *Tn5* binding bias to improve the accuracy of analysis: Early methods such as HINT-ATAC uses a position dependency model to infer the binding preferences [95], while TOBIAS creates a dinucleotide weight matrix to learn the underlying *Tn5* sequence bias from the observed *Tn5* cut sites from predefined peak calls [96]. More recently, deep learning methods such as ChromBPNet [97] and PRINT [98] applied fully convolutional structures with residual connections and dilated convolutions to facilitate bias correction at single-base-pair and single-cell resolution, as well as adapted to the complex signal patterns in multi-scale footprinting.

Experimentally, the ideal scenario is using ‘naked’ DNA (nucleosome- and histone-free DNA) as input to for transposase digestion to investigate *Tn5* insertion rates and sequence-binding bias. However, to my best understanding, there is currently no available study on ATAC-seq for naked DNA in mammalian cells. There were researches comparing the characteristics between naked DNA and nucleosome-binding DNA in yeast cells, but there was no study on chromatin accessibility. If possible, it is helpful to perform ATAC-seq on a sequence of mammalian DNA and on the synthetic naked DNA of the same sequence to directly investigate the differences of *Tn5* binding.

So far, there is no study taking account for the *Tn5* sequence binding bias in ATAC-seq peak calling. Indeed, peak calling has been considered as a pure statistical signal extraction problem precluding the consideration of the actual sequences. However, with the goal to reveal true accessibility in ATAC-seq peak calls, it is reasonable to optimise peak callers by adjusting to the sequence bias. For EpiCall, specifically, the mechanistic model itself over-simplifies the scenario that *Tn5* holds a constant insertion rate at ‘open’ and ‘closed’ region regardless of the sequences. In future work, EpiCall can be optimised as a sequence-dependent model, such as assigning different weights for single positions giving different surrounding sequences in the calculation of emission probabilities or hyperparameter search.

Improving hyperparameter optimisation for automatic fine tuning

The current settings for hyperparameters ρ_{open} and ρ_{closed} on all datasets are kept the same as the selected values from the grid search results of the GM12878 dataset. The selected hyperparameters produced satisfactory performance in most datasets, but there is still room for improvement, especially for bulk dataset constituting of heterogeneous cell population or multi-cell-type scATAC-seq datasets. In these cases, the ρ_{open} and ρ_{closed} at a single genomic position are not necessarily the same across all cells, which is against the hypothesis in the current model. It is theoretically beneficial and more biologically explicable to allow a varying base-wise insertion rates across cells. Performing light-weight fine tuning based on the original selected ρ_{open} and ρ_{closed} on selected candidate regions may also improve the performance in individual datasets.

In addition, it is necessary to investigate why the estimated ρ_{open} and ρ_{closed} has only a difference of two fold, as opposed to the algorithm assumption that $\rho_{open} \gg \rho_{closed}$. The same parameter grid search can be performed on other datasets of different cell types and sequencing depths to investigate the variations of the estimations across datasets. If the estimations vary significantly, the dependence between transposase insertion rates and data qualities should be considered in the model settings. On the contrary, if the estimations are about the same across datasets, it is worthwhile to discuss the biological plausibility in the definition of ρ_{open} and ρ_{closed} within the algorithm, and if increasing the number of hidden states produces more explicable values (Section 3.4.2).

The ratio between the estimated ρ_{open} and ρ_{closed} is also a quality control metric as an indicator of the noise level of the ATAC-seq dataset. For example, for the same cell/tissue population with a range of cell number, if the ratio between the estimated ρ_{open} and ρ_{closed} is positively related to the number of cells, it suggests ideally more samples should be collected to produce high-quality dataset to reveal the true patterns of the dataset. The ‘saturation point’ should be the lowest number of cells for the estimated ratio to stabilise.

The estimation of cell number also requires optimisation. It currently builds on the distribution of signal coverage in a selected candidate region, but preliminary experiments have shown instability in the estimation. In future updates, I will include the normalised signal coverage and average sequencing depth into the estimation.

3.5 Code and Data Availability

The implementation, tutorial and sample scripts of EpiCall are available at <https://github.com/luntergroup/EpiCall>. The data used in this Chapter are all taken from publicly available databases. For details of the data, see Appendix 1.

3.6 Methods

3.6.1 Data Preprocessing

Bulk ATAC-seq raw paired-end sequencing data in FASTQ was acquired from ENCODE and GEO, and aligned to genome build hg38 using Bowtie2 (version 2.4.4). Reads that are properly mapped, paired, having MAPQ>30 and are primary alignments were retained. For datasets without biological replicates, this is the input to call peaks. For datasets with more than one biological replicates, each replicate was pre-processed individually using the steps above and merged using `samtools merge` to a single BAM input to peak call. For scATAC-seq datasets, I took the 10X-style fragments file as the input to peak call if provided, otherwise I converted the filtered BAM alignment file to the fragments format to keep consistency between analysis. After peak calling, I remove peaks of any overlap with the ENCODE blacklists.

3.6.2 Peak Calling

MACS2

MACS2 [49] version 2.2.6 is used to peak call using `macs2 callpeak -format BAMPE -nolambda -keep-dup all` for inputs in BAM format, and `macs2`

`callpeak -format BEDPE -nolambda -keep-dup all` for inputs in 10X-style fragments format. The resulted narrowPeaks are merged (`bedtools merge`) and duplicates-removed to get the final peak set. For point-wise calculations for PR curves and AUC values as in Table 3.1 and Figure 3.3, a list of q -value cut-offs (`-q`: 0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00005, 0.00001) are used to produce different sets of peaks. All other analysis are performed using the result produced by the commonly-used q -value `-q 0.05`.

HMMRATAC

HMMRATAC [53] version 1.2.10 is used to call peaks for BAM inputs (merged across biological replicates). The Java program is set '`-window 2500000`' and run with '`-Xmx 64G`' to adapt to the memory of the computing system. Peak calls are reported in gappedPeak format and assigned with scores by default. I retained peaks with score ≥ 0 for downstream analysis. For point-wise calculations for PR curves and AUC values as in Table 3.1 and Figure 3.3, I subset the peaks by peak score cut-offs (5, 10, ..., 400) to produce different sets of peaks.

LanceOtron

LanceOtron [55] version 1.2.6 is used to call peaks from the coverage files. BigWig of RPKM normalised ATAC-seq coverage is generated from the individual BAM (if no biological replicates applicable) or the merged BAM (if given multiple biological replicates) using `bamCoverage` in `deepTools`, setting `-bs 1 -extendReads -normalizeUsing RPKM`. Peak calls are generated using `lanceotron callPeaks`, with score cut off at 0.5, followed by scoring for each peak using `lanceotron scoreBed`. For point-wise calculations for PR curves and AUC values as in Table 3.1 and Figure 3.3, I subset the peaks by peak score cutoffs (0.5, 0.6, 0.7, 0.8, 0.9, 0.99, 0.999, ... , 0.9999999) to produce different sets of peaks.

3.6.3 AUPRC measurement of ATAC-seq peak calls

I use two sets of references for AUPRC evaluation.

For the first set, I define the real positives by merging the active states (TSS Active, TSS Flank, TSS Flank Upstream, TSS Flank Downstream, Enhancer Active 1, Enhancer Active 2) and the real negatives by merging the repressive states (ZNF genes repeats, Heterochromatin, Repressed/Weak Repressed PolyComb, and Quiescent/Low) from the 18-state ChromHMM annotations [61–63]. ChromHMM is a hidden Markov Model that segments and annotates the human genome into disjoint chromatin states. The 18-state annotations are derived from integrative signals from 6 histone modification marks (H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3 and H3K36me3), producing specific annotations for each of the 127 reference epigenomes across different cell lines, primary cells and tissues in the ENCODE and Roadmap Epigenomics Project [6, 99].

For the second set, I define the real positives as the candidate CREs (cCREs) taken from the ENCODE SCREEN database [7], which was constructed by a look-up match between H4me3K1, H4me3K3, CTCF and DNase-seq signals. I define the real negatives the same as in the first set, but remove the regions which overlap with the cCREs defined as the real positives.

The details for creating subsets of peak calls to generate point-wise precision-recall curves and calculate AUPRC are introduced above in and Section 3.2.5 and Methods 3.6.2.

To calculate the point-wise precision and recall in order to plot the Precision-Recall curve and compute AUPRC, I take the similar approach as in the HMM-RATAC paper [53]: the point-wise true positives (TP) are defined as the base pairs overlapping between the called peaks and the real positives; False positives (FP) are defined as the base pairs overlapping between called peaks and the real negatives. Precision is calculated as $TP/(TP + FP)$; Recall (TPR) is calculated as TP/RP . F1 score is calculated as $2 * (Precision * Recall)/(Precision + Recall)$. The value of precision at the unreachable recall= 1 is defined as $(\text{the total number of real positives})/(\text{the total number of real positives} + \text{the total number of real negatives})$.

3.6.4 Motif enrichment analysis

To calculate the enrichment of each TF motif in ATAC-seq peaks, I use `findMotifsGenome.pl` in HOMER [86] with JASPAR2022 [60] as the known motif database for search. For each set of ATAC-seq peak calls, I randomly select 5 times the total number of GC-content-matched peaks (`-N 5`) from the genome as background control, set the search space for motif enrichment the exact size of each peak (`-size given`), and enable auto-normalization to reduce the sequence bias.

3.7 Acknowledgement

Figure 3.6 is generated using a powerful interactive visualisation platform, Multi-Locus-Viewer (MLV) [76], developed by my colleague Martin Sergeant. Thanks to Martin’s support, the public session of this analysis is available online as referred to in the figure caption.

4

EpiCall Iterative Peak Caller Facilitates Differential Regulatory Analysis Using scATAC-seq

Contents

4.1	Introduction	74
4.2	Results	75
4.2.1	Robustness of Peak Calling on Low Cell-number scATAC-seq Datasets	75
4.2.2	Differential Motif Enrichment Analysis for <i>Buenrostro2018</i> Dataset	76
4.2.3	Improved Cell Representation and Differential Regulatory Motifs analysis for 10X PBMC10K Multiomics Dataset	80
4.2.4	EpiCall Sensitively Recovers TF and Enhancer Signals from Proerythroblasts scATAC-seq Data	82
4.3	Discussion	86
4.3.1	Peak Calling for Rare Cells	86
4.4	Methods	87
4.4.1	scATAC-seq Data Pre-processing	87
4.4.2	Differential Motif Enrichment Analysis	88
4.4.3	PBMC10K Cell Type Annotations	88
4.4.4	Evaluation for Cell Clustering with Ground Truth	89
4.5	Acknowledgment	89

4.1 Introduction

Single-cell sorting and barcoding techniques have pushed the boundaries of genomic sequencing from aggregated signals of large cell population to the level of the single cell. Techniques such as scRNA-seq, scATAC-seq, and CITE-seq reveal the heterogeneity of the epigenomic, gene expression or proteomic profile across single cells, either as individual or multi-modal assays. As detailed in Section 1.4.2 and Figure 1.4.2), a number of downstream functional analyses of single-cell sequencing data, including cell clustering, differential motif enrichment, co-accessibility and trajectory inference often start from a *feature-by-cell* count matrix. In scATAC-seq, the *features* are defined as representative regions with enriched signals of chromatin accessibility, as either *fixed-size bins* across the genome or *peaks*. Compared with bulk ATAC-seq, scATAC-seq may suffer from additional challenges due to the signal of rare cell type being obscured in aggregated signals, such as the challenge of identifying peaks of rare cell types using the traditional ‘pseudo-bulked’ scATAC-seq peak calling which is based on the aggregated signals over cells. An alternative approach is *iterative peak calling*, which generates an initial set of peaks via pseudo-bulked peak calling, and clusters cells given the peak-by-cell count matrix, followed by peak calling from the pseudo-bulked signals over each cell cluster before merging into a common secondary peak set [37, 38]. Theoretically, this process can be performed iteratively until meeting pre-defined criteria such as a minimum cluster-to-cluster difference. The purpose of *iterative peak calling* is to enhance signals from rare cell types which may be obscured in the original pseudo-bulked peaks and reveal cluster-specific features [30, 39].

In this chapter, I will demonstrate the performance of EpiCall in revealing high-resolution functional heterogeneity for scATAC-seq using the *iterative peak calling* approach. I will first evaluate the robustness of calling peaks using EpiCall for different down-sampled cell number for a homogeneous scATAC-seq dataset, before presenting the differential motif enrichment results based on peak calls on *Buenrostro2018* (with ground-truth cell type annotations via FACS labels) and 10X

PBMC 5K multiome data (with downstream cell type annotations from scRNA-seq marker gene expression profiles). Lastly, I will demonstrate the sensitivity improvement in identifying regulator elements in an in-house pro-erythroblasts scATAC-seq population by using EpiCall.

4.2 Results

4.2.1 Robustness of Peak Calling on Low Cell-number scATAC-seq Datasets

Similar to bulk ATAC-seq, I evaluated the robustness of peak calling on scATAC-seq datasets of different sequencing quality. Instead of down-sampling by read number, for scATAC-seq datasets I down-sampled to different fraction of cell number to generate low-quality datasets of varied pseudo-bulked sequencing depth. I took the 10X 1:1 mixture of human GM12878 and mouse A20 scATAC-seq data (see Appendix 1 for data source) and extracted the human GM12878 aligned reads only, generating a scATAC-seq dataset of a relatively homogeneous cell line for the ease of non-bias down-sampling and evaluation. The data was further preprocessed to retain reads of properly paired, high mapping quality, and with associated barcodes that have passed the cell filtering in Cell-Ranger ATAC. The final raw dataset has 2302 cells with 149,307,841 reads. I randomly selected 70%, 50%, 30%, 10%, 5%, 3%, 1% and 0.5% of the total cells to generate down-sampled scATAC-seq datasets. Down-sampling is repeated three times with different seeds for each rate. Peak calling was performed using EpiCall with the optimised parameter settings, MACS2 with three different false discovery rates ($q=0.05$ default, 0.5 and 0.005), HMMRATAC with the default settings and LanceOtron with overall peak score cut-off at 0.5. For performance evaluation, I reported the Jaccard similarity index, recall and precision of the peak calls at base pairs with the same two different reference sets as in Section 3.3.4. Figure 4.1 shows the mean and standard deviations of each statistic at different down-sampling rates, with the x -axis showing the percentage of cells and cell number. Results show that the

precision of EpiCall is similar to MACS2 at $q=0.005$ and $q=0.05$ across down-sampled rates, while EpiCall consistently outperforms all three MACS2 settings in terms of recall and Jaccard similarity. In particular, EpiCall shows more significant advantages at cell number as low as 23 (1.0%) and 11 (0.5%), achieving higher recall and Jaccard similarity without sacrificing precision too much, while MACS2 at $q=0.005$ and $q=0.05$ achieves lower recall and MACS2 at $q=0.5$ drops significantly on precision. These suggest the robustness of EpiCall in peak calling scATAC-seq data on varied cell number and sequencing depth, which could be promising when applying to real-world scATAC-seq of rare cell types.

4.2.2 Differential Motif Enrichment Analysis for *Buenrostro2018* Dataset

Having confirmed the robustness and performance of EpiCall on scATAC-seq with low cell number, it is equally necessary to ask how well EpiCall is to accurately and sensitively identify differential epigenomic information from heterogeneous, quality-varied cell population, a common scenario in single-cell analysis. The first dataset I chose for this challenge is the *Buenrostro2018* dataset [39], which consists of the major hematopoietic lineages extracted from CD34+ bone marrow cells.

Hematopoiesis refers to the complex process of the differentiation and maturation of blood and immune cells (Figure 4.2 A). These multi-lineage heterogeneous cells are produced by different progenitor cells that differentiated from multipotent, self-renewing hematopoietic stem cells (HSCs) in bone marrow. Studies have shown that differentiation starts with HSCs which subsequently forms a hierarchy of lineage trajectories with increased lineage-specificity and decreased potency. Broadly speaking, HSCs first differentiate into multipotent progenitors (MPPs) which further generate lymphoid primed MPPs (LMPPs) and common myeloid progenitors (CMPs). The former produces lymphoid lineage progenitors, common lymphoid progenitors (CLPs, progenitors of adaptive immune cells B, T, and NK cells) and plasmacytoid dendritic cells (pDCs, which develop into peripheral pDCs), while the latter produces myeloid and erythroid lineage progenitors, such as

4. EpiCall Iterative Peak Caller Facilitates Differential Regulatory Analysis Using scATAC-seq

77

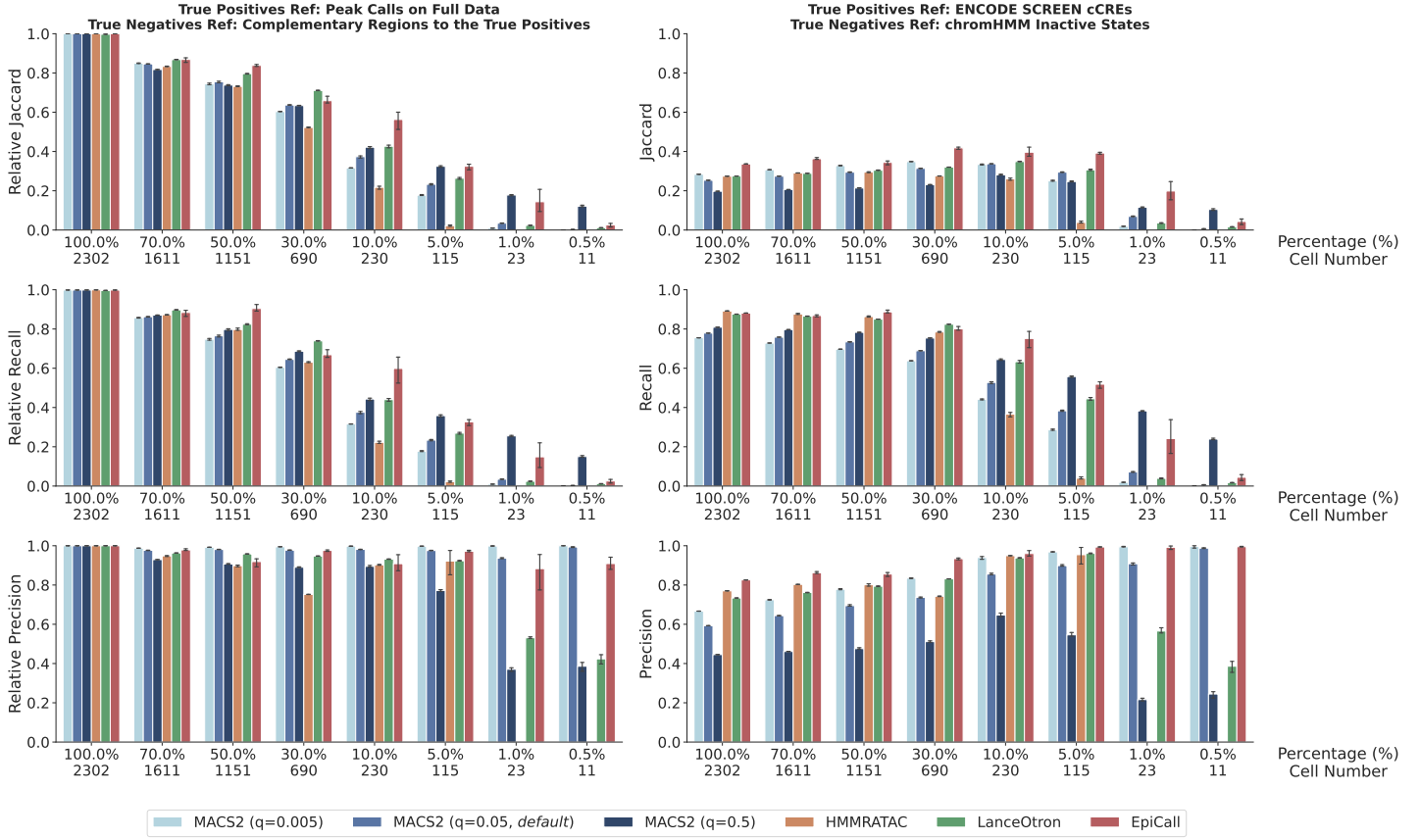


Figure 4.1: Peak calling performance on down-sampled 10X GM12878 scATAC-seq evaluated by ‘self comparison’ (left) and ‘general performance’ (right). The x-axis labels the percentage of cells and cell number at each down-sampling rate. Notice that HMMRATAC fails to peak call at down-sampling rates 1.0% and 0.5%. Colors represent the three peak callers, in which MACS2 is performed at three different false discovery rates (q values) 0.005, 0.05 (default) and 0.5. For the reference regions used for evaluation, the left plot serves as a ‘self comparison’ by setting the true positive regions as the peak calls in respective to each peak calling setting on the raw data without down-sampling, and the true negative regions as the complement of the real positives. The statistics in self comparison shows the relative Jaccard similarity, recall and precision in respective to the raw peak calls. The right plot, in complementary, evaluates the ‘general performance’ of peak calls the same using real positives as the cell-type specific ENCODE SCREEN cCREs and real negatives as the 18-state ChromHMM negative states.

erythroid megakaryocyte-erythroid progenitors (MEPs, progenitors of erythroblasts and megakaryocytes, which finally produce red blood cells and platelets through maturation, respectively). Note that both LMPPs and CMPs can differentiate into granulocyte macrophage progenitors (GMPs, progenitors of innate immune cells monocytes and mDC granulocytes).

The characterisation and isolation of major hematopoietic lineages are originally dependent on immunophenotypical features, such as using fluorescence activated

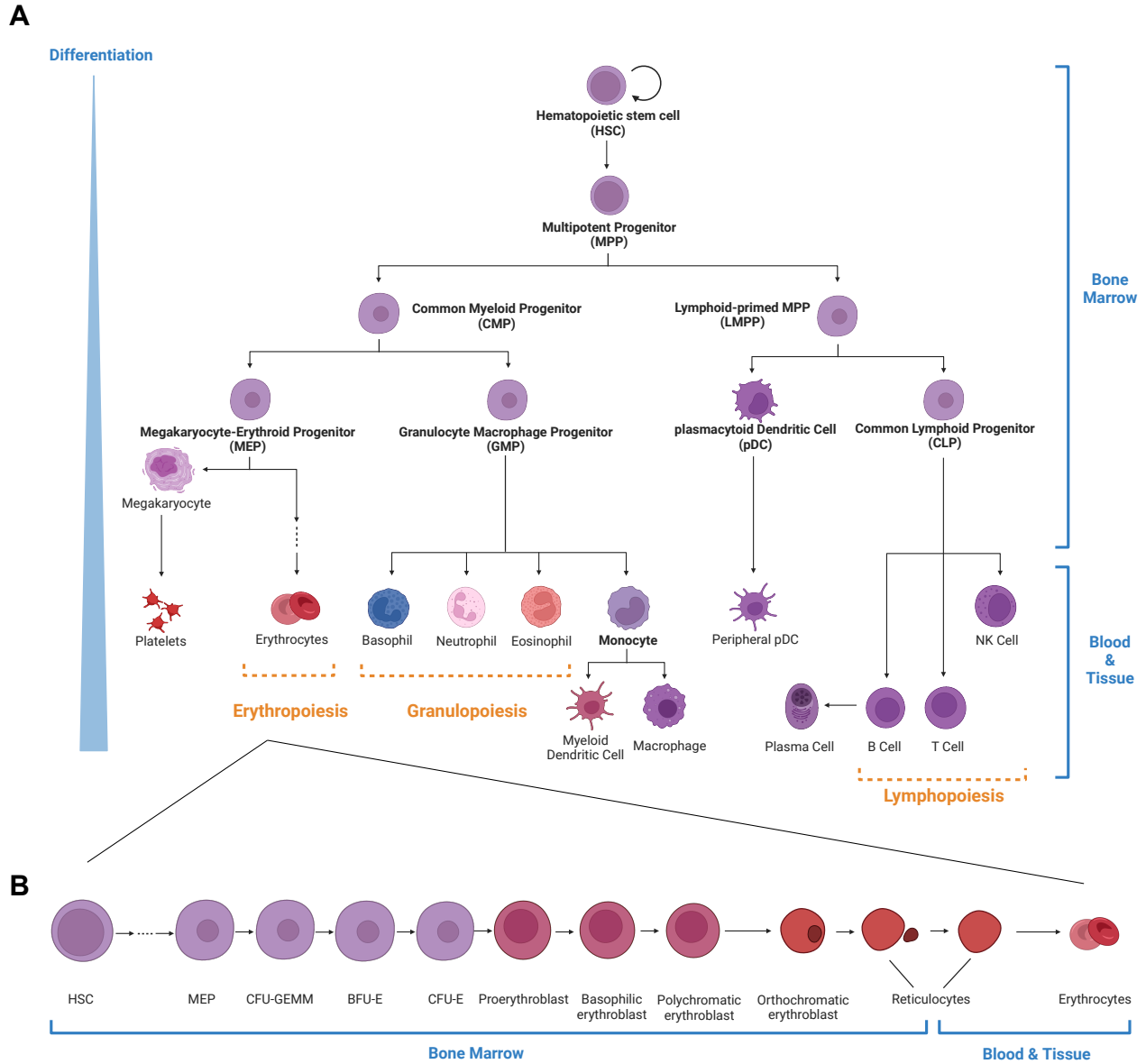


Figure 4.2: Schematic view of human hematopoiesis with emphasis on erythropoiesis. **A.** Schematic view of human hematopoiesis showing differentiation levels and marking erythropoiesis, granulopoiesis and lymphopoiesis. Cell types used in the *Buenrostro2018* dataset are highlighted in bold. Based on template <https://app.biorender.com/illustrations/645d807b47410b63d2910e01> by Akiko Iwasaki. **B.** Schematic view of human erythropoiesis based on [100]. Both charts are created with Biorender.com.

cell sorting (FACS) [30, 101]. Integrated with ATAC-seq techniques, Buenrostro *et.al.* [39] isolated human CD34+ bone marrow cells from 6 donors into 9 cell types (HSCs, MPPs, LMPPs, CMPs, CLPs, pDCs, GMPs, MEPs and Monocytes) and the rest as unknown (UNKs) according to their cell surface protein markers

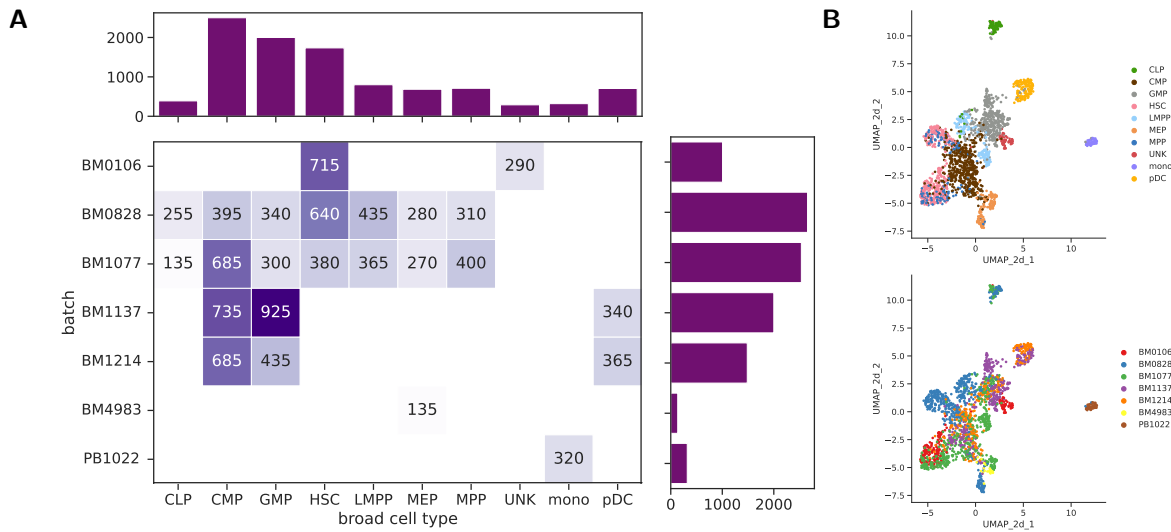


Figure 4.3: Summary of the *Buenrostro2018* dataset. **A.** The number of cells in each cell type, batch and joint combinations. **B.** Cell embeddings on 2D UMAP using ArchR TileMatrix (500bp fixed-size window region vs.cell count matrix), coloured with broad cell annotations given by the ground-truth FACS labels (top) and batches (bottom). HSC - hematopoietic stem cell, MPP - multipotent progenitor, LMPP - lymphoid-primed MPP, CLP - common lymphoid progenitor, pDC - plasmacytoid dendritic cell, CMP - common myeloid progenitor, GMP - granulocyte macrophage progenitor, MEP - megakaryocyte-erythroid progenitor, Mono - monocyte and UNK - unknown.

[39]. Each cell is labeled and collectively digested with *Tn5* and sequenced to create the scATAC-seq libraries, generating in total 2034 cells with 383,889,489 reads after quality control. The FACS labels are considered the gold-standard cell types in downstream analysis. As an essential supplement to transcriptomics and phenotypical features, the *Buenrostro2018* dataset reveals the heterogeneity in the epigenome and regulatory mechanisms during hematopoiesis of not only the known cell types, but also between sub-populations of HSCs, MPPs and GMPs that were obscured previously. Figure 4.3A summarizes the number of cells of each donor, FACS cell type, and the joint pair.

Studies on the complex regulatory dynamics of *Buenrostro2018* have shown the motif variability of certain TFs responsible for the lineage-specific differentiation in hematopoiesis. Using this dataset as a benchmark, for each peak caller, I performed two types of peak calling: the first calls on the entire combined BAM which merges over all cells; and the second performs a one-step iterative peak calling by calling on each separated BAM of the aggregated alignments over the same cell type from

the FACS label before merging the peaks into a union peak set.

I then performed differential motif enrichment analysis using chromVAR [46] based on each set of peaks, investigating the significance and variance of motifs across each single cell as well as across the ground-truth labeled cell types. chromVAR produces a GC% bias-corrected *variability* for each motif as a measure of the differential level across cells; a *z*-score for each motif and cell as a measure of gain or loss of motif activity (see Section 4.4.2 and [46] for method details).

The results show that all cell-type-specific peak sets recapitulated known trajectories in terms of motif enrichment, a proxy for the activity of the associated transcription factor. Among the six peak sets analyzed, these trajectories are best resolved when using the peak set produced by EpiCall using the iterative peak call strategy (Figure 4.4A deep red tracks). Specifically, EpiCall peak calls shows the greatest range of motif enrichment and depletion *z* scores across the cell types, providing the clearest contrast between motifs within each cell type and even in each individual cell, as visualized in Figure 4.4B.

The representative motifs chosen for Figure 4.4 demonstrates the advantage of EpiCall for analyzing motif enrichment. CEBPB, a known regulator of monocyte development, shows the highest activity in monocytes and depletion in most other cell types; GATA1, a key regulator of the erythroid lineage, is most active in MEPs and shows activity in its precursor CMP while being low in lymphoid lineages; and ID3, a regulator of lymphoid development, presents highest activity in pDC, followed by CLP and LMPP, while being lowest in MEP and monocytes.

4.2.3 Improved Cell Representation and Differential Regulatory Motifs analysis for 10X PBMC10K Multiomics Dataset

Following the analysis of *Buenrostro2018* [39], I continued to explore the performance of EpiCall in more sophisticated datasets. One example is 10X PBMC 10K multiome from a healthy donor (see Appendix 1 for data source), which provides a joint profiling of scRNA-seq and scATAC-seq for each single cell. After preprocessing and quality control, the remaining dataset consists of 10,989 cells. I performed the

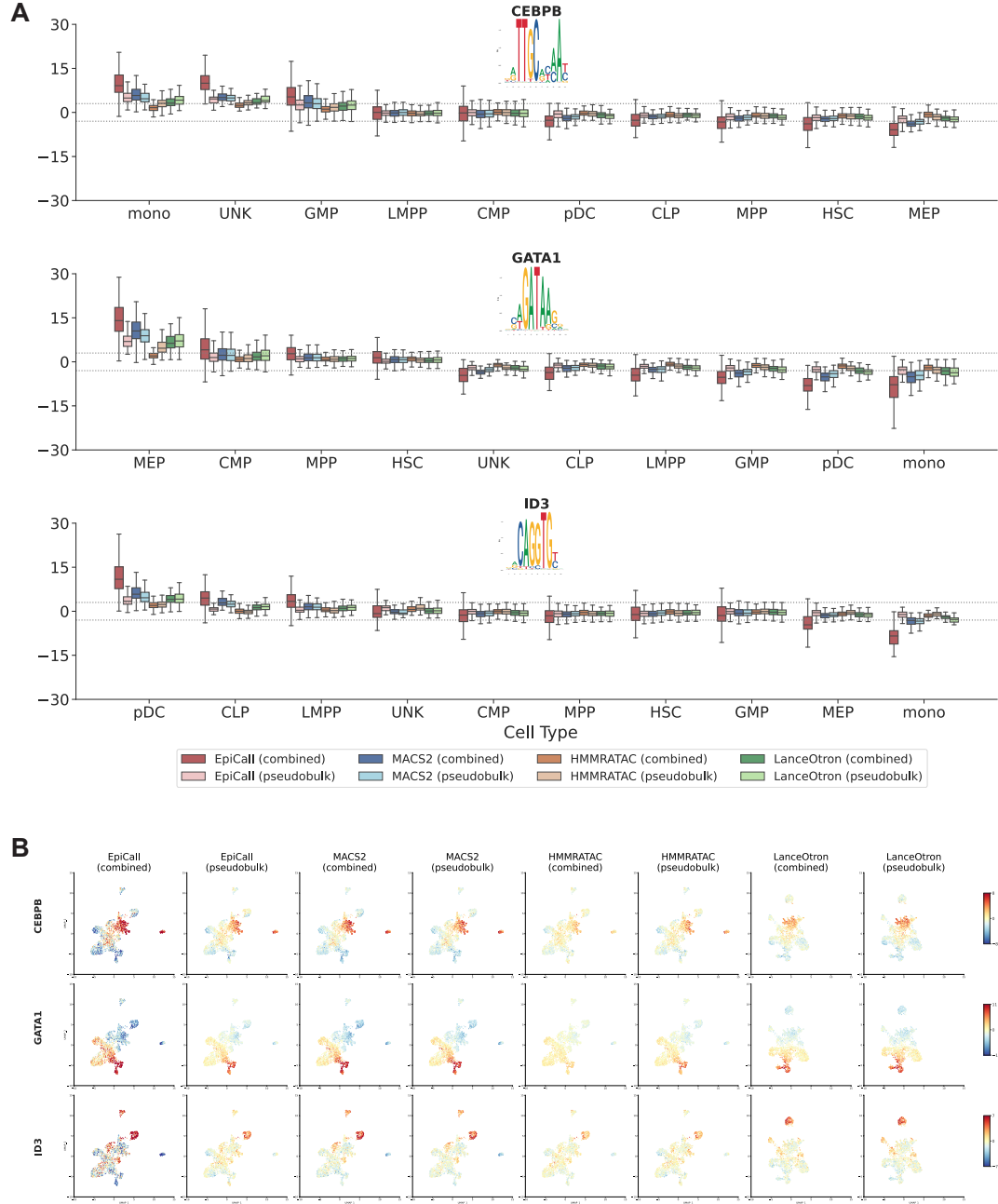


Figure 4.4: chromVAR differential motif enrichment of representative motifs on *Buenrostro2018*. **A.** Boxplots showing the z-scores of differential enrichment of CEBPB, GATA1 and ID3 motifs, grouped by cell type and colored by peak calling strategies used to construct *peak vs. cell* matrix. For each motif, the order of the cell types on the x-axis corresponds to the averaged enrichment of the motif with respect to each cell type. The boxes extend from the first quartile (Q1) to the third quartile (Q3) of the data, with a line showing the median. The whiskers extend from the boxes by 1.5x the inter-quartile range (IQR). grey dotted lines mark z-score = ±3; **B.** Scatter plots on 2D UMAP coordinates showing per-cell z-scores of the same peak calling strategies (columns) and motifs (rows) as in (A). The color gradient corresponds to the z-scores of TF enrichment, in which for each motif, the threshold of the colorbar on the side is set to 0.025 and 0.975 quantile across all cells and all peak calling strategies of that motif.

same differential motif analysis using chromVAR using the count matrix created with peak calls produced by EpiCall and MACS2, respectively (Figure 4.5).

Consistent with the previous analysis, all peak calling methods have correctly captured the trajectories across cells via the predicted activity z -scores of the five lineage/cell-type representative motifs from the JASPAR2020 database: CEBPA, a key regulator of monocyte development; SPIB, a lymphoid-specific enhancer; POU2F2, a B-cell-specific transcription factor; TCF7, an important regulator in the differentiation of T cells; and GATA1, an essential TF in erythroid and megakaryocyte differentiation. Again, across the five representative motifs, the peak set made by EpiCall using the iterative peak calling strategy produce the best inferred differential trajectories (Figure 4.5C) as the strongest enrichment in the expected cell types.

Since this is a multi-omics dataset, Seurat offers cell type annotations for each single cell based on their scRNA-seq gene expression profile via a transfer-label approach in reference with a large pre-annotated PBMC dataset [29] (Methods 4.4.1). Using the annotations as ground-truth labels, I evaluated the cell clustering performance based on the four different sets of peak calls by measuring statistics Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI) and the Homogeneity (Methods 4.4.4). While all four peak callers produce similar Homogeneity, EpiCall using the iterative peak calling strategy shows the highest ARI and AMI, indicating a potential superiority in distinguishing cell types from complex mixture of cells (Figure 4.5B).

4.2.4 EpiCall Sensitively Recovers TF and Enhancer Signals from Proerythroblasts scATAC-seq Data

Erythropoiesis belongs to a part of hematopoiesis, referring the differentiation and maturation of erythrocytes (i.e. red blood cells) (Figure 4.2B). During erythropoiesis, the epigenomic regulators are gradually turned off, and cells at different stages can be harvested to generate stage-wise cell population. During in-house experiments, my colleagues differentiated CD34+ hematopoietic stem cells *in-vitro* from a healthy

4. EpiCall Iterative Peak Caller Facilitates Differential Regulatory Analysis Using scATAC-seq

83

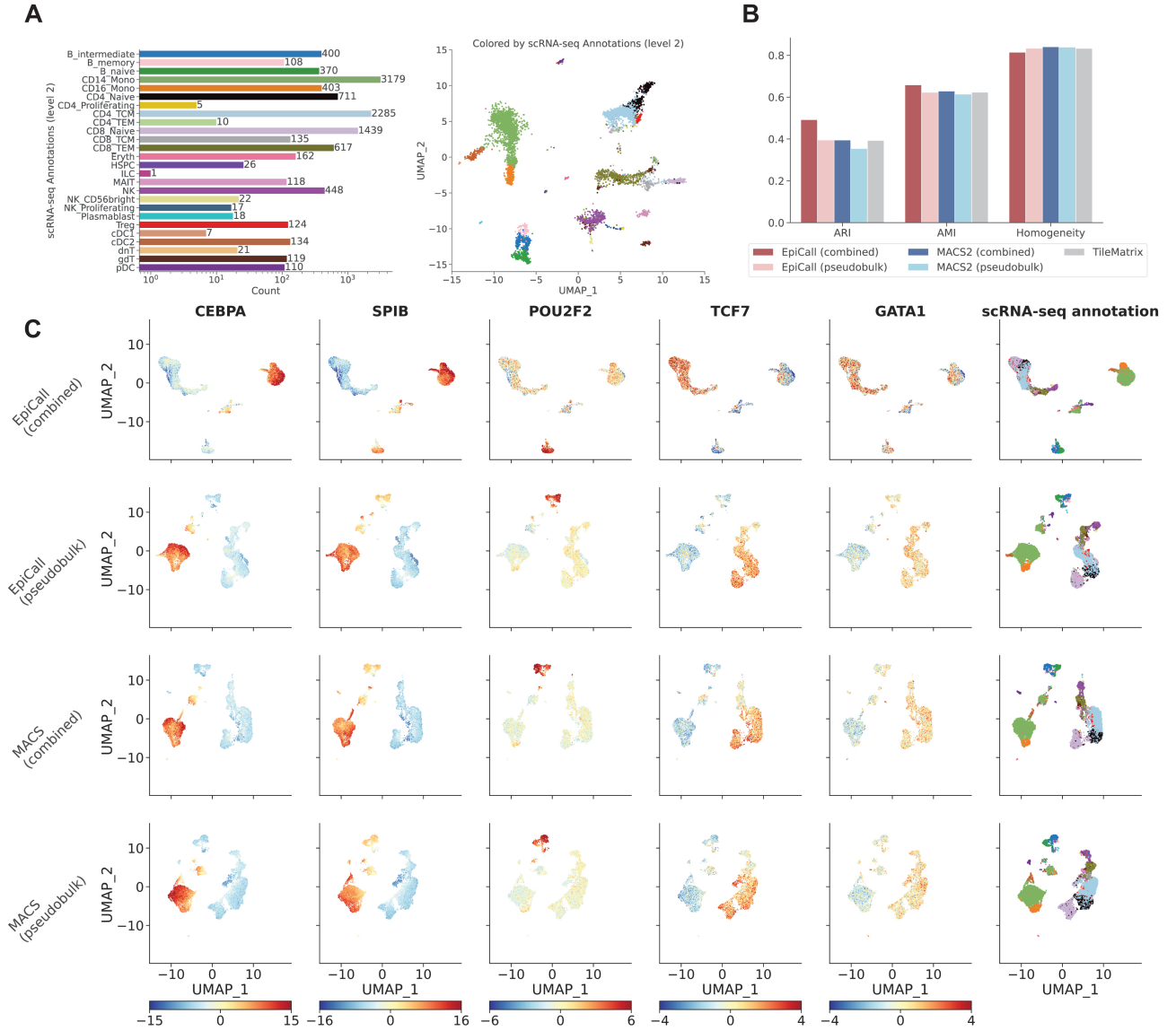


Figure 4.5: Cell type annotations, cell clustering performance and chromVAR differential motif enrichment of representative motifs on 10X PBMC 10K multiome.

A. Summary statistics of scRNA-seq annotations. Single cells are annotated based on the scRNA-seq profile using Azimuth level2 cell types. Barplots showing the number of cells of each annotated cell type; scatter plots showing the 2D UMAP embeddings of the scRNA-seq gene-by-cell count matrix with the same color scheme as the barplots. **B.** Clustering performance measured using ARI, AMI and Homogeneity (Methods 4.4.4) between the scATAC-seq clusters and the scRNA-seq cell type annotations, colored by the sets of regions used for clustering: peak calls from four peak calling methods, respectively, and 500bp-window tile matrix. **C.** Scatter plots on 2D UMAP coordinates showing per-cell z-scores of the same peak calling strategies (rows) and selected representative motifs (columns 1-5) along with the cell type annotations by paired scRNA-seq (last column). The color gradient corresponds to the z-scores of TF enrichment, in which for each motif, the threshold of the colorbar at the bottom is set to 0.025 and 0.975 quantile across all cells and all peak calling strategies of that motif. The UMAP coordinates are computed based on the peak-by-cell count matrix constructed by individual sets of peak calls. The color scheme of the scRNA-seq annotations are kept the same as in **A**.

male human donor and conducted 10X scATAC-seq on proerythroblasts collected at Day 10 after differentiation, right before the expulsion of the nuclei [102].

I performed peak calling on the pseudo-bulked proerythroblasts scATAC-seq reads using EpiCall, MACS2, HMMRATAC and LanceOtron, and navigated around two highly complex and challenging gene loci of interest for comparison.

Figure 4.6 focuses on the KLF1 gene locus, a known active regulatory region in proerythroblasts. Here I demonstrate the signal coverage of scATAC-seq data, peak calls, as well as TF and Histone ChIP-seq around the locus. Specifically, I marked the signals of all fragments (Figure 4.6 first row, red shade) as well as the filtered fragments of size ≤ 150 bp showing the nucleosome-free signals (Figure 4.6, first row, red line). The nucleosome-free scATAC-seq fragments reveal high-resolution signal enrichment in six non-overlapping regions, of which five correspond to the binding signals of CTCF, GATA1 and NFE2 in the same cell line (Figure 4.6, purple tracks), which are indistinguishable in the unfiltered signals. As peak calling is performed on the raw fragments without size filtering, it is challenging to identify high-resolution signals from the background. The results of EpiCall stand out by sensitively revealing and distinguishing separated peaks corresponding to the enrichment of the nucleosome-free scATAC-seq signals and the representative TF bindings. Although EpiCall ‘over-called’ certain peaks without clear evidence of nucleosome-free enrichment, the color depth of the peaks, corresponding to the peak score representing their significance, are consistent with whether they correlate with rich nucleosome-free signals. Among the other three peak callers, LanceOtron and HMMRATAC are poor at distinguishing peaks in high resolution, while MACS2 fails to identify peaks hidden in the low-resolution signals. The yellow highlighted region (Figure 4.6, highlighted region) shows a potential weak but true peak, with clear evidence of CTCF binding but atypical shape of scATAC-seq signals, is successfully called as an individual peak by EpiCall, missed by MACS2, and is not distinguishable from neighbouring peaks in LanceOtron and HMMRATAC.

Similarly, Figure 4.7 shows the promoter region of SLC25A37 gene locus, an actively locus encoding a mitochondrial iron transporter specifically responsible

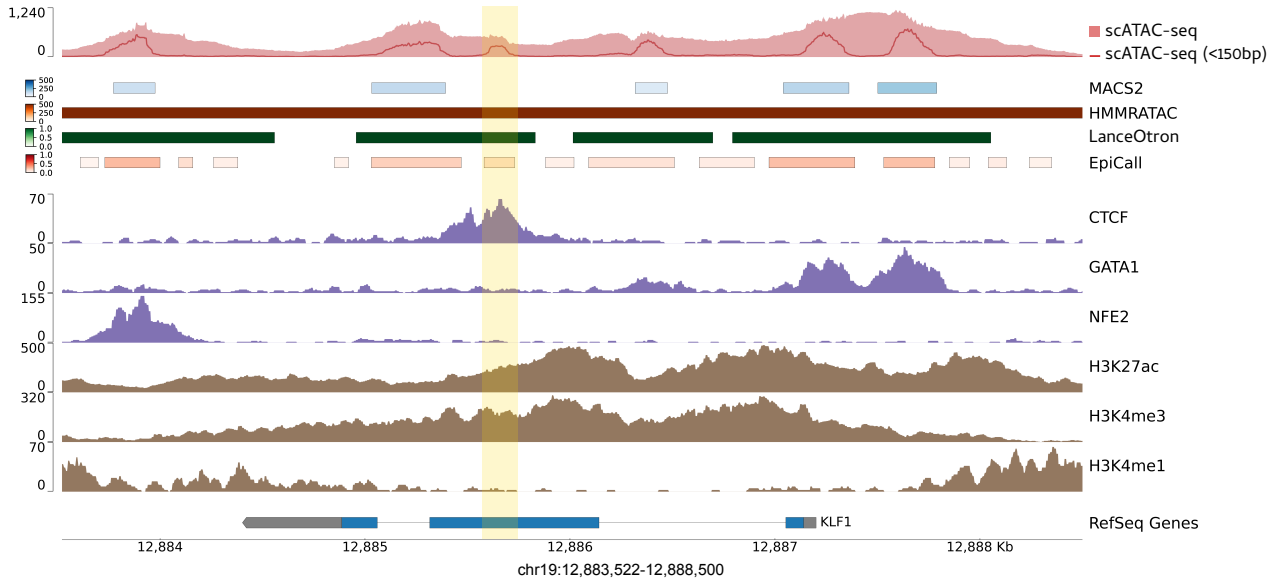


Figure 4.6: scATAC-seq proerythroblast at KLF1 locus, showing correspondence with ChIP-seq coverage of TF and Histone modifications. From top to bottom: Track 1: aggregated scATAC-seq coverage of all fragments (red shade) and fragments ≤ 150 bp (red line). Track 2-5: peak calls produced by MACS2, HMMRATAC, LanceOtron and EpiCall on the entire aggregated scATAC-seq fragments, with color depth corresponding to peak scores of each peak caller. The color range is shown on the left for each track. Track 6-11: ChIP-seq signal coverage of CTCF, GATA1, NFE2, H3K27ac, H3K4me3 and H3K4me1. Track 12: RefSeq showing KLF1 locus.

for iron import in proerythroblast cells and erythroids [103]. Again, I show the signal coverage of scATAC-seq data, peak calls, and ChIP-seq of TF and Histone modifications. The signal enrichment of the nucleosome-free scATAC-seq data (Figure 4.7 first row, red line) reveals two distinct cCREs, each corresponding to the binding of CTCF and GATA1, respectively (Figure 4.7, highlighted region). Among the peak callers, only EpiCall successfully recognises and distinguishes the cCREs and assigns with high peak scores, while MACS2, HMMRATAC and LanceOtron fail to separate the peaks. In addition, the unfiltered scATAC-seq signals contain noisy signals throughout the gene body of the SLC25A37 locus with no supporting evidence in neither TF nor open-chromatin Histone marks, which confuses HMMRATAC and LanceOtron that produce a number of ‘over-called’ peak calls likely to be false positives, whereas MACS2 and EpiCall both overcome the challenge.

In both examples above, EpiCall is the only peak caller that both identifies and distinguishes all potential true peaks with evidence of signal enrichment

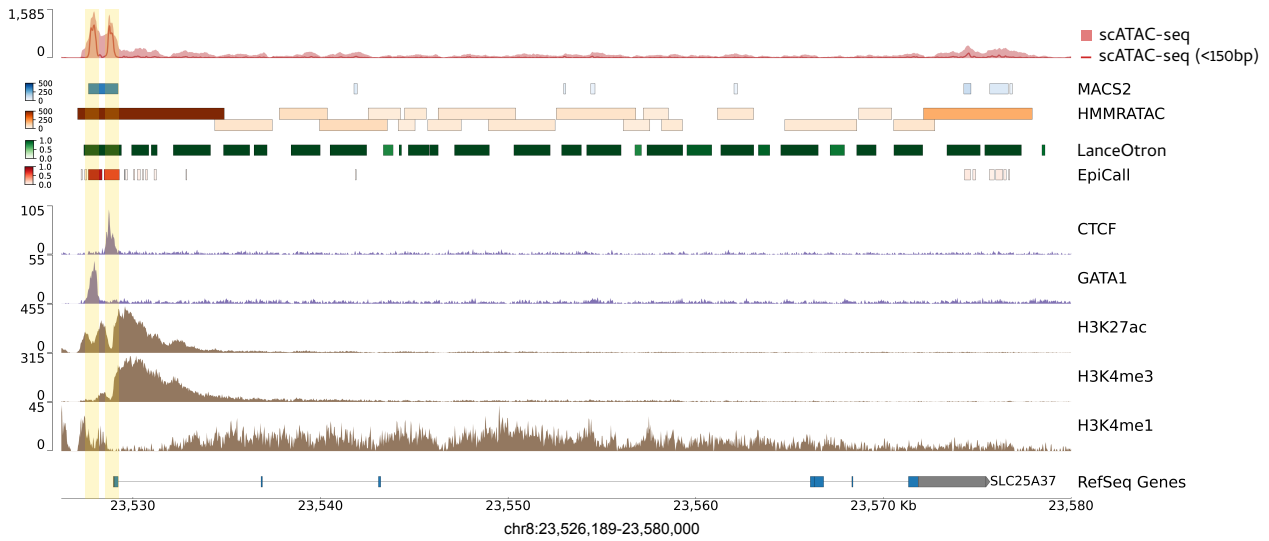


Figure 4.7: scATAC-seq proerythroblast around SLC25A37 locus, showing correspondence to ChIP-seq coverage of TF and Histone modifications. From top to bottom: Track 1: aggregated scATAC-seq coverage of all fragments (red shade) and fragments ≤ 150 bp (red line). Track 2-5: peak calls produced by MACS2, HMMRATAC, LanceOtron and EpiCall on the entire aggregated scATAC-seq fragments, with color depth corresponding to peak scores of each peak caller. The color range is shown on the left for each track. Track 6-10: ChIP-seq signal coverage of CTCF, GATA1, H3K27ac, H3K4me3 and H3K4me1. Track 11: RefSeq showing SLC25A37 locus.

in nucleosome-free scATAC-seq, CTCF and cell-type representative TF binding, and histone marks of active regulatory elements. These results demonstrate the advantage of EpiCall in balancing the sensitivity and resolution in peak calling on individual locus level, consistent with the genome-wide analysis and locus-based examples on bulk ATAC-seq data in Chapter 2.

4.3 Discussion

4.3.1 Peak Calling for Rare Cells

Signal extraction on rare cells has been challenging across bulk and single-cell sequencing. The high quality of single-cell sequencing has increased the resolution of signal detection greatly, but several difficulties still exist in the peak calling for rare cells:

1. The absolute number of cells can be low. Insufficient cell number (i.e. lower than 5 cells) will lead to inevitable high noise levels that diminish their signals

in pseudo-bulked peak calling [34]. However, in biological tissues, certain types of cells can be rare inherently, such as HSPC and erythroblasts in PBMC, early malignant cells in tumor micro-environment, *etc.*.

2. The relative biological difference between the rare cells and the neighbouring abundant cells can be small. This is most critical when requiring high-level annotations of cells within similar cells.
3. Potential novel cell types are difficult to annotate.

As introduced in Section 1.7.2, deep learning approaches have been using sparse signal and corresponding low-quality peak calls to predict high-quality peak calls, achieving reasonable predictions in both same cell types and across different but similar cell types [81, 104]. However, these supervised learning models assume homogeneity through the down-sampling process in model training, which may be sub-optimal to solve the challenges listed above. Focusing on the peak calling algorithms, the results in this Chapter have demonstrated EpiCall’s capability of producing robust peaks given lower cell number than other peak callers. In addition, a simple one-step iterative peak calling strategy accompanied with EpiCall has improved the differential functional analysis. This iterative strategy can be potentially extended to a multi-iterations, for which the Jaccard similarity can be used to measure the difference of the union of peak calls between the current and last iteration. When the difference is smaller than a pre-defined threshold, iterations can be terminated that produce a final peak set.

4.4 Methods

4.4.1 scATAC-seq Data Pre-processing

The scATAC-seq data are pre-processed using a comprehensive analytical tool, ArchR [36]. First, the scATAC-seq fragments are processed to generate a fixed-size-window based tile matrix using ArchR. Next, cells are dimension reduced using iterative Latent Semantic Indexing (iLSI) [29, 105], clustered using Seurat’s

graph-based approach [106] and embedded onto 2D dimension space using UMAP. To allow the comparison between peak sets produced by different peak calls, iLSI, clustering and UMAP embeddings are performed on each *peak vs. cell* count matrix in addition to the common tile matrix.

4.4.2 Differential Motif Enrichment Analysis

For *Buenrostro2018* and 10X PBMC 10K multiome datasets, I used chromVAR [46] to calculate the differential motif activity across single cells. Based on each *peak vs. cell* count matrix, a *motif vs. cell* count matrix is derived representing the enrichment significance of motifs across each cell. For each motif and cell, chromVAR first computes a *raw variability*, the difference between the total number of fragments that map to peaks containing the motif and the expected number of fragments (based on the average of all cells). To account for the GC content biases in the aggregation, they created for each motif a set of *background* genomic regions which has an equal number and are matched for GC% and average accessibility. The raw variability for these background regions are used to compute a bias-corrected deviation (i.e. *variability*) and the *z*-score for each motif and cell, providing a differential measure of the gain or loss of accessibility of a motif relative to the average cell profile.

4.4.3 PBMC10K Cell Type Annotations

The identities of cells are annotated based their paired scRNA-seq gene expression profile using a reference-based label transfer approach by Seurat [107]. To annotate PBMC cells, Seurat uses a pre-annotated large reference dataset, Azimuth ¹, which enables automatic transfer labeling to single cells of any query PBMC dataset [107]. The level 2 Azimuth cell types were used for annotations and the ground truth to evaluate cell clustering.

¹<https://azimuth.hubmapconsortium.org/references/>

4.4.4 Evaluation for Cell Clustering with Ground Truth

To evaluate cell clustering results (Y_{pred}) for datasets provided with ground-truth labels (Y_{true}), I use three well-established metrics: adjusted Rand index (ARI), adjusted Mutual Information (AMI), and Homogeneity.

The Rand index (RI) [108] is a measure of the similarity between two clusters. ARI then accounts for chance for RI:

$$\text{ARI}(Y_{pred}, Y_{true}) = \frac{\text{RI}(Y_{pred}, Y_{true}) - \mathbb{E}(\text{RI}(Y_{pred}, Y_{true}))}{\max(\text{RI}(Y_{pred}, Y_{true})) - \mathbb{E}(\text{RI}(Y_{pred}, Y_{true}))}$$

in which $\mathbb{E}$ is the expected Rand index.

The Mutual Information (MI) is a measure of the mutual dependence between the two random variables. It quantifies the ‘amount of information’ obtained about one variable by observing the other: $\text{MI}(Y_{pred}, Y_{true}) = D_{KL}(P(Y_{pred}, Y_{true}) \mid P(Y_{pred}) \otimes P(Y_{true}))$ ². Again, AMI [109] accounts for the chance for MI:

$$\text{AMI}(Y_{pred}, Y_{true}) = \frac{\text{MI}(Y_{pred}, Y_{true}) - \mathbb{E}(\text{MI}(Y_{pred}, Y_{true}))}{\max(\text{H}(Y_{pred}), \text{H}(Y_{true})) - \mathbb{E}(\text{MI}(Y_{pred}, Y_{true}))}$$

in which $\text{H}(Y)$ is the entropy associated with the clustering Y .

Homogeneity [110] is used to measure how similar the samples in a cluster are:

$$h = 1 - \frac{\text{H}(Y_{true} \mid Y_{pred})}{\text{H}(Y_{true})}$$

in which $\mathbb{H}(Y_{true})$ is the entropy of the true labels, and $\text{H}(Y_{true} \mid \text{H}(Y_{pred}))$ is the conditional entropy of the true labels given the predicted labels.

All three metrics score between 0 and 1, and higher values suggest better clustering performance.

4.5 Acknowledgment

The Proerythroblasts ATAC-seq alignment data (BAM) and ChIP-seq signal coverage (BigWig) used in Section 4.2.4 are generated in in-house experiments

² D_{KL} is the Kullback-Leibler Divergence; $P(Y_{pred})$, $P(Y_{true})$ are marginal distributions; $P(Y_{pred}, Y_{true})$ is the joint distribution; and $P(Y_{pred}) \otimes P(Y_{true})$ is the outer product distribution which assigns probability ($P_{Y_{true}}(y_{true})$)($P_{Y_{pred}}(y_{pred})$) to each (y_{pred}, y_{true}) .

by colleagues from a normal human donor using 10X Genomics Chromium™ i7 Multiplex Kit N Set A (1000084) [102] and referred to me by Emine Ravza Gur. The alignments data and metadata of *Buenrostro2018*, the alignments data of 10X GM12878 + A20 scATAC-seq and 10X 10K PBMC multiome are taken from public databases listed in Appendix 1.

5

Conclusion

The increasing amount, diversity and resolution of NGS data has been driving large-scale statistical modeling, functional analysis and biological interpretation towards the ultimate goal of deciphering gene regulation. The best models involve the deep understanding of the data characteristics and close integration with domain knowledge. This thesis is built on the fact that the different fragment sizes contribute differently to the signal composition and functional information within ATAC-seq data. The following chapters demonstrated the efficient use of the fragment size information for two different purposes: i) constructing an scATAC-seq analysis pipeline for high-resolution functional variant prioritisation (AVOCATO); and ii) developing a novel ATAC-seq-specific peak caller (EpiCall), which provides sensitive and robust peak calls on data of varying quality and facilitates the analysis of differential functional genomics across single cells in both abundant and rare cell populations.

In this chapter, I first summarise the key contents of this thesis, followed by discussing the possible improvements for evaluating ATAC-seq peak calling and scATAC-seq dimension reduction. Then I envision the possibility of setting up systematic benchmarking and datasets for scATAC-seq analysis paradigms in the near future.

5.1 Summary

5.1.1 High-resolution scATAC-seq analysis aids the prioritisation of functional variants

In Chapter 2, I introduced a collaboration project, AVOCATO, an analytical and visualisation platform for scATAC-seq. By employing a simple but powerful fragment filtering strategy with several algorithmic optimisation in data processing, AVOCATO was successfully applied to prioritise functional genetic variants with biological interpretations from high-resolution scATAC-seq data analysis.

5.1.2 A mechanistic model as a robust and sensitive ATAC-seq peak caller

In Chapters 3 and 4, I developed EpiCall, an ATAC-seq specific peak calling algorithm that shows superior performance over three representative peak callers: MACS2, HMMRATAC and LanceOtron. The performance was evaluated on several metrics: i) A rough assessment of the peak calling quality: calculating the precision and recall between the peak calls and cell-type-specific ‘active’ and ‘inactive’ functional chromatin states annotated by ChromHMM; ii) Enrichment analysis of functional genomics, including quantitative measurement using HOMER and qualitative enrichment of cell-type specific TF binding (CTCF and enhancers) and histone modifications (H3K27ac, H3K4me3, H3K4me1) in peak regions; and iii) Robustness analysis of peak calls in downsampled data. In addition, leveraging an iterative peak calling strategy with EpiCall in scATAC-seq data facilitated more significant differential motif enrichment and more accurate identification for rare cell types. These evaluations comprehensively demonstrated the superior performance and its great potential to become a new, robust ATAC-seq peak caller adaptable to varying data quality.

Indeed, EpiCall is not the first algorithm incorporating ATAC-seq fragment size in modeling peak calling, as HMMRATAC utilises the information in defining the hidden states in the HMM construction (though in a more direct way). Here we want to address that through the results on various datasets demonstrated

in this thesis, the *mechanistic model* which models the formation of ATAC-seq fragments provides the basis for a natural integration of the fragment size and coverage information, facilitating both an improved performance in robust peak calling and biological explicability.

It is also a promising direction to combine EpiCall with machine learning models for peak calling. With the ability to produce sensitive peak calls, EpiCall can be used to identify candidate peaks rigorously, followed by applying machine learning models for peak filtering, such as using LanceOtron [55] to assign peak scores and filter the peaks with different cut-offs depending on the usage.

Despite of the confidence for EpiCall, there are still several aspects of the model can be improved, which I have discussed extensively in Discussion 3.4 in Chapter 3.

5.2 Better functional annotations in need for ATAC-seq peak calling evaluation

Among the three criteria I used for evaluating EpiCall (Section 5.1.2), measuring the overlap between peak calls and known ‘gold-standard’ functional annotations is a common practice widely used in recent peak callers HMMRATAC [53], RCL [54] and many tutorials. In this thesis, the gold standard we took were the 18-state ChromHMM cell-type specific ‘active’ (true positives, grouping states 1_TssA, 2_TssFlnk, 3_TssFlnkU, 4_TssFlnkD, 9_EnhA1 and 10_EnhA2) and ‘inactive’ states (true negatives, grouping states 12_ZNF/Rpts, 13_Het, 16_ReprPC, 17_ReprPCWk and 18_Quies). However, ChromHMM and many other annotations (e.g. SegWay [65] and SCREEN cCREs [7]) are mostly derived from models based on bulk ChIP-seq assays for Histone modifications in the ENCODE databases. These annotations are not ideal due to the varying data qualities, missing assays, and the inherently broader signal patterns than chromatin accessibility. In fact, ATAC-seq signals show non-negligible enrichment in regions outside of the common definition of ‘active’ states (TssA, TssFlnk, TssFlnkU, TssFlnkD, EnhA1, EnhA2) and even in the ‘inactive’ states (ZNF/Rpts, Het, ReprPC, ReprPCWk, Quies). For example, in Figure 5.1, a larger fraction of peak calls overlap with the ‘weak

enhancer' state (EnhWk) than the 'active enhancer' states (EnhA1 and EnhA2), but only less than 10% of the 'weak enhancer' regions are recovered in the peak calls. Therefore, peaks locate neither in the 'true positives' and 'true negatives' are overlooked in evaluation. In addition, though it could improve the accuracy and resolution of functional annotations by including ChIP-seq assays of TF binding, it is impractical to experimentally conduct ChIP-seq for all known TF motifs or analytically take into account all TF ChIP-seq assays across cell types for modeling.

The existing studies on profiling *cis*-regulatory elements are mostly limited to the non-coding genome. Indeed, the coding regions takes only around 2% of the genome and a even smaller percentage of ATAC-seq peaks overlap with the annotated gene bodies (e.g. Figure 5.1, the overlap between ATAC-seq peaks in GM12878 and ChromHMM states). However, studies in thyroid cancer cells have identified enhancers within gene bodies, mostly in introns, which operates further distances than the promoters, interacts with chromatin loops, and strongly correlates with the gene transcription and protein expression levels [111]. Observations on in-house datasets (Figure 4.6) also show enriched chromatin accessibility within gene bodies. These all suggest more investigations into gene body-associated in functional regulations, especially on recent, high-quality datasets.

To construct better functional annotations suitable for peak calls produced from high resolution ATAC-seq and scATAC-seq data, it is necessary to build more precise functional annotations based on the NGS data generated in recent years, such as the single-cell multi-omic cell atlas. Similar to the label-transfer hierarchical cell type annotations in Azimuth [107], it would be beneficial to build and perform functional annotation frameworks from high-quality reference scATAC-seq dataset for both non-coding genome and gene bodies. In addition, involving assays capturing genome-wide chromatin interaction, such as HI-C and high-resolution 3C can improve the functional annotations for cell-type specific super enhancers and promoter hubs [112].

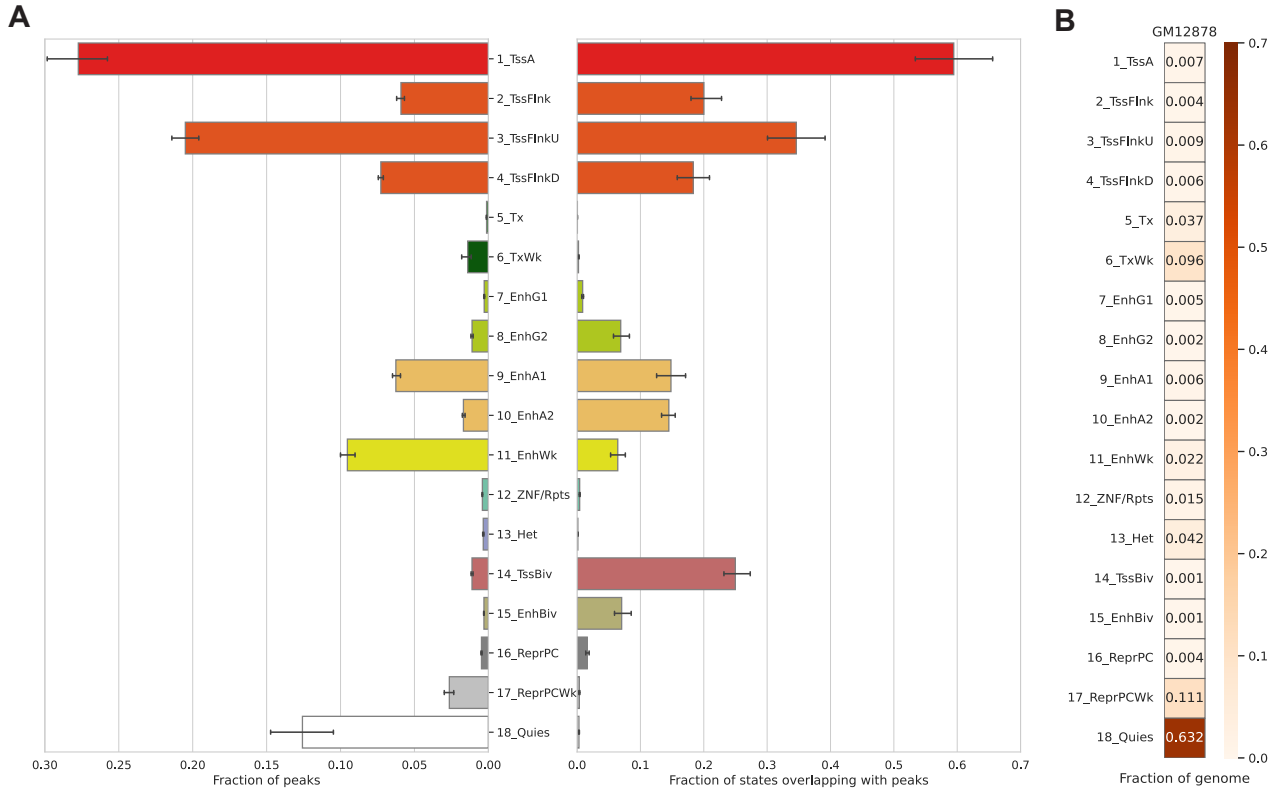


Figure 5.1: ChromHMM annotations are sub-optimal for benchmarking ATAC-seq peak calls. Example summary plots on GM12878 cell line: **A.** Barplots showing the fraction of base pairs of ATAC-seq peak calls overlapping with each 18-state ChromHMM state (left) and the fraction of base pairs of each ChromHMM state overlapping with ATAC-seq peak calls. ATAC-seq peak calling is performed on the GSE47753 GM12878 50K merged alignment data (see Appendix 1 for data source) using MACS2, HMMRATAC, LanceOtron and EpiCall with default parameters. For each state, the barplot shows the average across the four sets of peak calls with error bars showing the standard deviations. **B.** Heatmap showing the fraction of genome taking by each ChromHMM state.

5.3 Single cell representation learning: dimension reduction and cell state characterisation

scATAC-seq analysis is a multi-stage process and can involve various functional interpretations depending on individual use cases. Throughout the analysis, dimension reduction is an essential step to reduce the data complexity of the high-dimensional region-by-cell count matrix before downstream analysis. Due to even higher drop-out rate than scRNA-seq, statistical methods with superior performance on sparse data are applied to produce a low-dimensional embedding or modeling of scATAC-seq. Existing algorithms include Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (LDA) [42, 43], Jaccard with Non-Negative Matrix Factorisation (NMF),

Variational Auto-Encoders (VAE) [44] and Convolutional Neural Networks (CNN). Many are adopted from natural language processing (NLP).

LSI is a linear transformation approach adopted from NLP. The binarized region-by-cell count matrix is first transformed using term-frequency inverse-document-frequency (TF-IDF) for normalising read depth and enhancing regions appearing less frequently across cells (based on the assumption that they are more likely to be informative). The last step is to apply Singular Vector Decomposition (SVD) to the TF-IDF matrix to get the lower-dimension embeddings. LSI is employed by pipelines such as Cusanovich2018 [40], Signac [113] and 10X Cell Ranger ATAC[29]. Modifications including iterative LSI [29] and estimated LSI [36] have the potential to further reduce noise and improve sensitivity.

LDA is a probabilistic topic modeling technique used in NLP to identify latent semantic topics in a collection of documents, in which each document is represented as a mixture of topics, and each topic is a multivariate distribution over words. The entire corpus is then constructed as a generative model word by word by sampling the topic mixture of a document and then the word distribution of the topic [42, 43]. In scATAC-seq, LDA was adopted by cisTopic [41] to simultaneously co-optimize the classification of peak regions (words) into regulatory topics and the clustering of cells (documents) based on their topic contributions. The derived topic loading in the cell-by-topic distributions serves as a low-dimensional representation for the single cells. Besides, the inferred topic distributions offer full biological explanations in which the cell-by-topic distributions refer to the stable states of cells and the topic-by-peak distributions indicating co-regulatory elements such as enhancers.

Being widely applied in representation learning for high-dimensional data, VAE has also been naturally used in single-cell sequencing such as scATAC-seq. It learns a latent representation of the input as a variational distribution (encoder) and reconstructing the input space from the lower-dimensional space (decoder), and is optimised through minimizing the reconstruction error and the minimizing the Kullback-Leibler (KL) divergence between the learned approximate and the true posteriors. As a generative model, it also aids to the denoise of the data.

In scATAC-seq, SCALE [114] combined VAE with the Gaussian Mixture Model (GMM) to represent scATAC-seq data for denoising, imputation and cell clustering.

Recently, scBasset [45] extends Basset [47], a deep CNN framework predicting chromatin accessibility from fixed-size window of DNA sequences across cell lines from bulk ATAC-seq, to scATAC-seq, creating a more expressive sequence-dependent architecture for cell representation. In scBasset, the first convolutional layer learns sequential features such as TF motifs as in the original Basset, while the multiple tasks in the final layer of Basset become single cells with the weights learning cell embeddings. scBasset achieves performance improvement as well as interpretability in cell clustering, denoising, TF motif activity inference and multi-modal integration with scRNA-seq.

To the best of our understanding, there is no single best practice for scATAC-seq representation learning and dimension reduction. Depending on individual use cases, the evaluation of dimension reduction involves but is not limited to cell clustering, differential accessibility characterisation, motif analysis and correspondence to paired/unpaired assays of transcriptomics or proteomics.

In the age of multi-omics and the increasing demand of integrative analysis, graph neural networks (GNNs) are promising candidates for multi-modal single-cell representation learning. GNNs have been applied in scRNA-seq denoising [69, 70] and scATAC-seq TF regulatory network inference [115]. In GNNs, cells are regarded as nodes with similarities as edges, in which the similarities can be characterised by the distance between the sequencing profiles of individual cells. Therefore, different modalities can be used to construct different distance metrics before feeding into the same neural network architecture that integrates modalities in learning.

5.4 Towards systematic benchmarking and setting up scATAC-seq analysis paradigms

Continuous emergence of novel protocols in scATAC-seq analysis have been rapidly enriching the field, while a systematic, reproducible and hands-on benchmarking plat-

form remains incomplete. This brings obstacles for both developers to make convincing arguments on the model performance and users to assess methods and pipelines.

In 2019, Chen *et.al.* presented a benchmarking platform for the major scATAC-seq pipelines up to then [34]. The platform integrates the pre-processing steps of each method and measures the performance of cell clustering as the main criteria. Since then, several end-to-end softwares and pipelines, such as MAESTRO [116], scATAC-pro [117] and ArchR [36], have been trying to employ best practices at each analytical stage and efficiently incorporate into a unified workflow. Recently, De Pop *et.al.* benchmarked eight methods across 47 experiments on the human PBMCs based on quality control, cell clustering, and the calling of differentially accessible region (DAR), followed by developing a unified, modular pipeline, PUMATAC, for data pre-processing and allowing potential future updates [118]. In addition to method development, task-driven challenges have formed into enriching platforms of solving cutting-edge single-cell interpretation problems along with setting up benchmarking frameworks: *The Open Problems in Single-cell Analysis* tries to address and tackle the grand challenges in single-cell data science [119] with mathematically formalized tasks, open challenges and state-of-the-art benchmarking.

These practices set solid foundation to bring in a unified benchmarking platform as a potential community-based project, which potentially holds the following characteristics:

1. Create modular functionalities with maintenance and hands-on incorporation of new methods. The core team, assuming experts in the field, need to define the major stages to include in the platform, and work with software engineers to set up unified input/output paradigms with well-documented protocols for incorporation. From my personal experience, it will be ideal to build infrastructure using workflow management system, such as Snakemake[77], for reproducibility and scalability.
2. Provide a group of representative datasets used for method benchmarking. For simulated datasets to meet varying requirement on sequencing depth,

cell type mixtures and sparsity, the platform needs to set up a standard and parameterised simulation procedure. For real datasets, well-constructed decentralized projects such as the Human Cell Atlas can be a starting database.

3. Connect existing projects. There are currently several open-source projects focusing on different challenges in single-cell sequencing, whether large or small, overlap with the perspective of the benchmarking platform. Bringing connections between the communities will most efficiently gather domain expertise and facilitate system analysis, for example, benefiting the topic of integrative analysis of single-cell multi-omics.

scATAC-seq, since its emergence, has caught the interest of biologists, statisticians and computer scientists, who have been continuously driven the research and method development in a vivid, community-empowered environment. I believe the effort towards building a comprehensive benchmarking platform with well-structured datasets is at the right time point, and will make best use of the assay and accelerate the healthy growth of the field.

Appendices

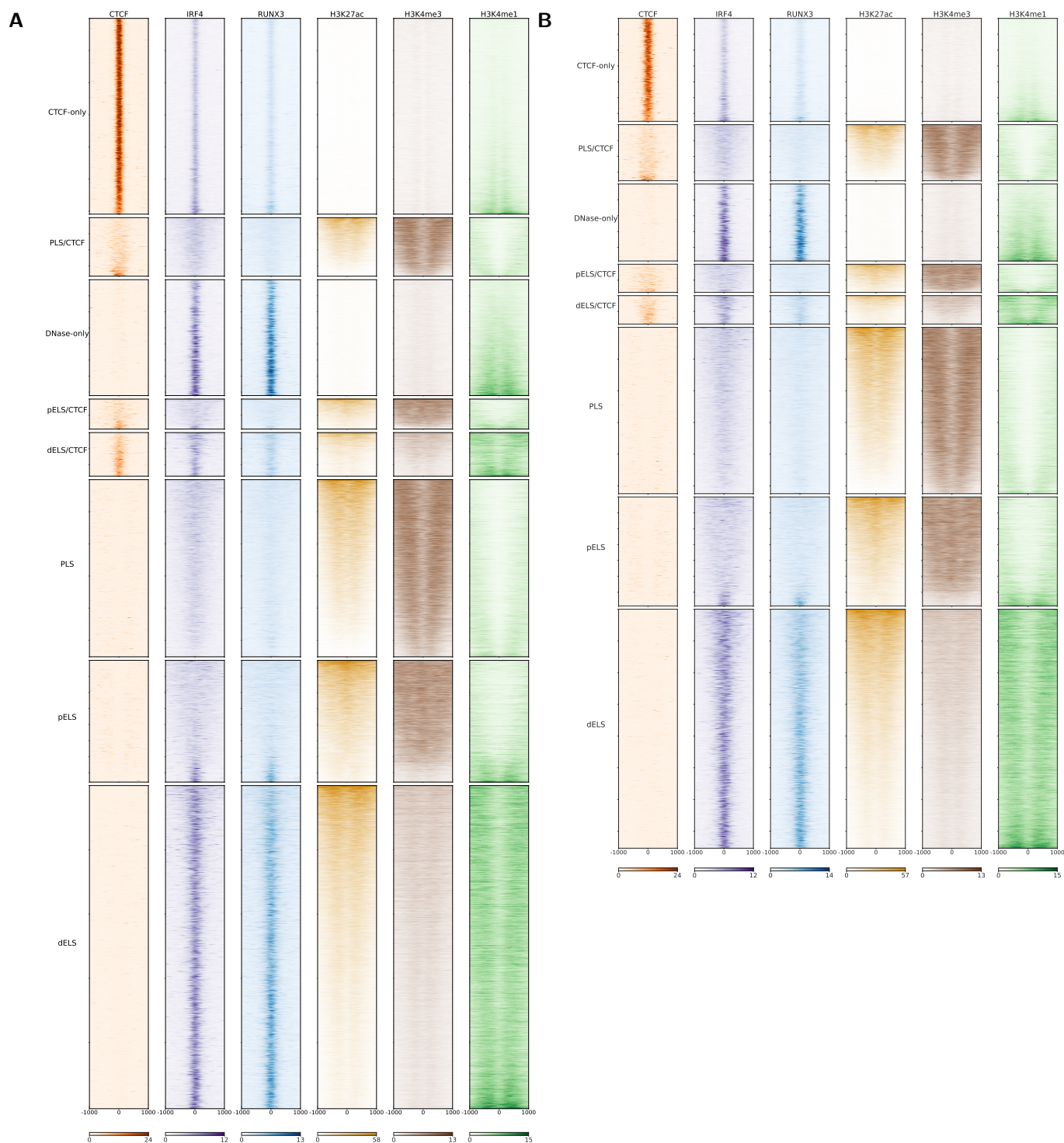


Figure 2: Signals of functional genomics surrounding the center of EpiCall and LanceOtron peak calls on GM12878 50k dataset. Pileup heatmap showing the ChIP-seq signals of CTCF, IRF4, RUNX3, H3K27ac, H3K4me3 and H3K4me1 surrounding the center of EpiCall peak calls (**A**) and LanceOtron peak calls (**B**). Each column is grouped by ENCODE SCREEN cCREs annotations, from top to bottom: CTCF-only, PLS/CTCF, DNase-only, pELS/CTCF, dELS/CTCF, PLS, pELS and dELS.

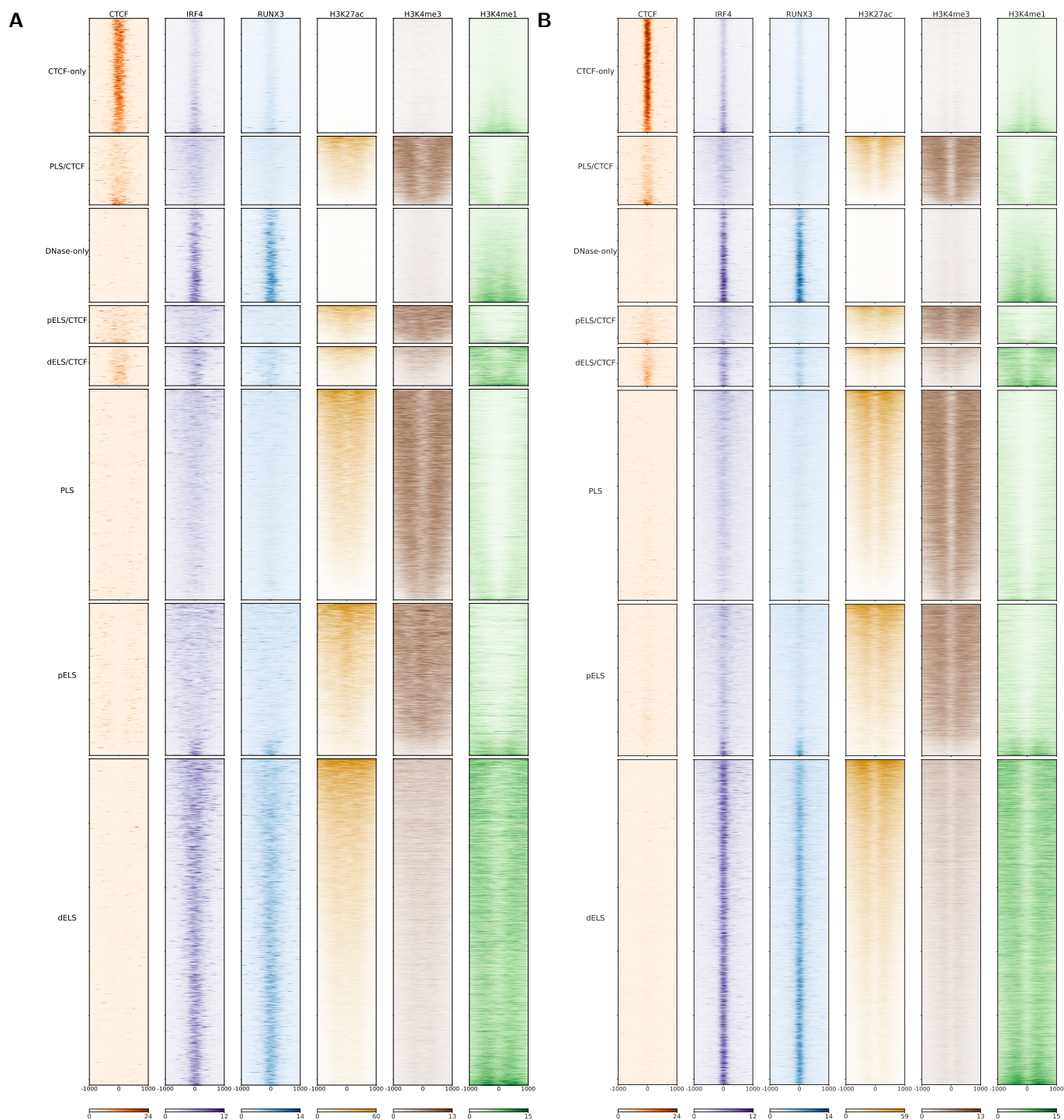


Figure 3: Signals of functional genomics surrounding the center and summit of MACS2 peak calls on GM12878 50k dataset. Pileup heatmap showing the ChIP-seq signals of CTCF, IRF4, RUNX3, H3K27ac, H3K4me3 and H3K4me1 surrounding the center (**A**) and summit (**B**) of MACS2 peak calls. Each column is grouped by ENCODE SCREEN cCREs annotations, from top to bottom: CTCF-only, PLS/CTCF, DNase-only, pELS/CTCF, dELS/CTCF, PLS, pELS and dELS.

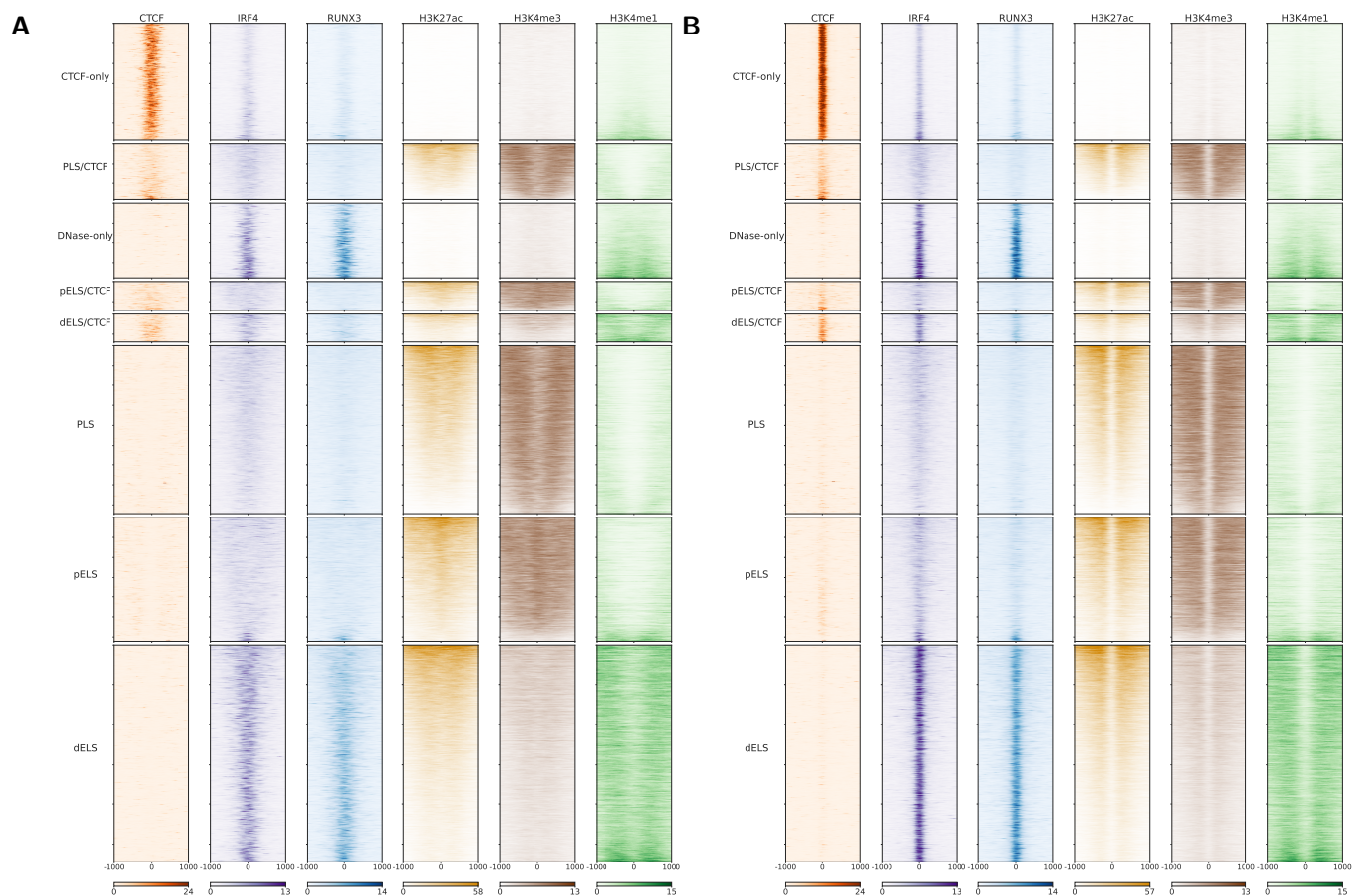


Figure 4: Signals of functional genomics surrounding the center and summit of HMMRATAC peak calls on GM12878 50k dataset. Pileup heatmap showing the ChIP-seq signals of CTCF, IRF4, RUNX3, H3K27ac, H3K4me3 and H3K4me1 surrounding the center (**A**) and summit (**B**) of HMMRATAC peak calls. Each column is grouped by ENCODE SCREEN cCREs annotations, from top to bottom: CTCF-only, PLS/CTCF, DNase-only, pELS/CTCF, dELS/CTCF, PLS, pELS and dELS.

Cell line/type	Type	GEO Session ID	ENCODE ID	File format	Notes
GM12878	Bulk ATAC-seq	GSM1155957	-	BAM	Merging 3 replicates
GM12878	CTCF ChIP-seq	GSM935611	ENCSR000DZN	BigWig	-
GM12878	IRF4 ChIP-seq	GSM803390	ENCSR000BGY	BigWig	-
GM12878	RUNX3 ChIP-seq	GSM1010893	ENCSR000BRI	BigWig	-
GM12878	H3K27ac ChIP-seq	GSM733771	ENCSR000AKC	BigWig	-
GM12878	H3K4me1 ChIP-seq	GSM733772	ENCSR000AKF	BigWig	-
GM12878	H3K4me3 ChIP-seq	GSE95899	ENCSR057BWO	BigWig	-
GM12878 ^a	10X scATAC-seq	-	-	BAM	-
K562	Bulk ATAC-seq	GSE170214	ENCSR483RKN	BAM	Merging 2 replicates
K562	CTCF ChIP-seq	GSM935407	ENCSR000EGM	BigWig	-
K562	TA1 ChIP-seq	GSE91490	ENCSR106FRG	BigWig	-
K562	GATA1 ChIP-seq	GSM1003608	ENCSR000EFT	BigWig	-
K562	NFE2 ChIP-seq	GSE95875	ENCSR000FCC	BigWig	-
K562	H3K27ac ChIP-seq	GSM733656	ENCSR000AKP	BigWig	-
K562	H3K4me1 ChIP-seq	GSM733692	ENCSR000AKS	BigWig	-
K562	H3K4me3 ChIP-seq	GSE96303	ENCSR668LDD	BigWig	-
HepG2	Bulk ATAC-seq	GSE170012	ENCSR042AWH	BAM	Merging 2 replicates
HepG2	CTCF ChIP-seq	GSM803486	ENCSR000BIE	BigWig	-
HepG2	HNF4A ChIP-seq	GSM803460	ENCSR000BLF	BigWig	-
HepG2	FOXA1 ChIP-seq	GSM803432	ENCSR000BMO	BigWig	-
HepG2	H3K27ac ChIP-seq	GSM733743	ENCSR000AMO	BigWig	-
HepG2	H3K4me1 ChIP-seq	GSM798321	ENCSR000APV	BigWig	-
HepG2	H3K4me3 ChIP-seq	GSE96248	ENCSR575RRX	BigWig	-
IMR-90	Bulk ATAC-seq	GSE169767	ENCSR200OML	BAM	Merging 2 replicates
Panc1	Bulk ATAC-seq	GSE170824	ENCSR591PIX	BAM	Merging 2 replicates
Human CD34+ Bone Marrow	FACS-sorted scATAC-seq	GSE96769	-	BAM	-
Human PBMC ^b	10X PBMC 3k multiomics	-	-	BAM	-
Human PBMC ^c	10X PBMC 10k multiomics	-	-	BAM; GEX count matrix	-
Proerythroblasts	scATAC-seq	-	-	BAM	in-house data
Proerythroblasts	CTCF ChIP-seq	-	-	BAM	in-house data
Proerythroblasts	GATA1 ChIP-seq	-	-	BAM	in-house data
Proerythroblasts	NFE2 ChIP-seq	-	-	BAM	in-house data
Proerythroblasts	H3K27ac ChIP-seq	-	-	BAM	in-house data
Proerythroblasts	H3K4me1 ChIP-seq	-	-	BAM	in-house data
Proerythroblasts	H3K4me3 ChIP-seq	-	-	BAM	in-house data
PhyloP ^d	Evolutionary conservation	-	-	BigWig	-
18-state chromHMM annotations ^e	Functional annotations	-	-	BED	-

Table 1: Source of data used in this thesis

^a[https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_v1_hgmm_5k_5k_1:1_mixture_of_fresh_frozen_human_\(GM12878\)_and_mouse_\(A20\)_cells](https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_v1_hgmm_5k_5k_1:1_mixture_of_fresh_frozen_human_(GM12878)_and_mouse_(A20)_cells).
Extract only the human GM12878 cells to use in this thesis.

^b<https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-3-k-1-standard-2>

^c<https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2>

^d<http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phyloP100way/>

^ehttps://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/core_K27ac/jointModel/final/

	State	Description
1	TssA	Active TSS
2	TssFlnk	Flanking TSS
3	TssFlnkU	Flanking TSS Upstream
4	TssFlnkD	Flanking TSS Downstream
5	Tx	Strong transcription
6	TxWk	Weak transcription
7	EnhG1	Genic enhancer1
8	EnhG2	Genic enhancer2
9	EnhA1	Active Enhancer 1
10	EnhA2	Active Enhancer 2
11	EnhWk	Weak Enhancer
12	ZNF/Rpts	ZNF genes & repeats
13	Het	Heterochromatin
14	TssBiv	Bivalent/Poised TSS
15	EnhBiv	Bivalent Enhancer
16	ReprPC	Repressed PolyComb
17	ReprPCWk	Weak Repressed PolyComb
18	Quies	Quiescent/Low

Table 2: 18-state ChromHMM Annotations

References

- [1] Chris M. Rands et al. “8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage”. In: *PLOS Genetics* 10.7 (July 2014), pp. 1–12. DOI: 10.1371/journal.pgen.1004525.
- [2] Francis S Collins, Michael Morgan, and Aristides Patrinos. “The Human Genome Project: lessons from large-scale biology”. In: *Science* 300.5617 (2003), pp. 286–290.
- [3] Feng Yan et al. “From reads to insight: a hitchhiker’s guide to ATAC-seq data analysis”. In: *Genome biology* 21 (2020), pp. 1–16.
- [4] Yin Chen et al. “A constitutive super-enhancer: homologous region 3 of Bombyx mori nucleopolyhedrovirus”. In: *Biochemical and biophysical research communications* 318.4 (2004), pp. 1039–1044.
- [5] Denes Hnisz et al. “Super-enhancers in the control of cell identity and disease”. In: *Cell* 155.4 (2013), pp. 934–947.
- [6] Anshul Kundaje et al. “An integrated encyclopedia of DNA elements in the human genome”. eng. In: *Nature (London)* 489.7414 (2012), pp. 57–74.
- [7] Yunhai Luo et al. “New developments on the Encyclopedia of DNA Elements (ENCODE) data portal”. eng. In: *Nucleic acids research* 48.D1 (2020), pp. D882–D889.
- [8] Bradley E Bernstein et al. “The NIH roadmap epigenomics mapping consortium”. In: *Nature biotechnology* 28.10 (2010), pp. 1045–1048.
- [9] Casey E Romanoski et al. “Roadmap for regulation”. In: *Nature* 518.7539 (2015), pp. 314–316.
- [10] Roadmap Epigenomics Consortium Integrative analysis coordination Kundaje Anshul 1 2 3 Meuleman Wouter 1 2 Ernst Jason 1 2 4 Bilenky Misha 5, Scientific program management Chadwick Lisa H. 53, and Principal investigators Bernstein Bradley E. 2 26 42 Costello Joseph F. 14 Ecker Joseph R. 9 Hirst Martin 5 18 Meissner Alexander 2 6 Milosavljevic Aleksandar 7 Ren Bing 8 13 Stamatoyannopoulos John A. 10 Wang Ting 21 Kellis Manolis 1 2. “Integrative analysis of 111 reference human epigenomes”. In: *Nature* 518.7539 (2015), pp. 317–330.
- [11] Susan Fairley et al. “The International Genome Sample Resource (IGSR) collection of open human genomic variation resources”. In: *Nucleic acids research* 48.D1 (2020), pp. D941–D947.
- [12] Annalisa Buniello et al. “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019”. In: *Nucleic acids research* 47.D1 (2019), pp. D1005–D1012.

- [13] Elliot Sollis et al. “The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource”. In: *Nucleic Acids Research* 51.D1 (2023), pp. D977–D985.
- [14] A Balsalobre and J Drouin. “Pioneer factors as master regulators of the epigenome and cell fate”. In: *Nat Rev Mol Cell Biol* 7.23 (2022), pp. 449–464.
- [15] Kazuhiro Maeshima, Satoru Ide, and Michael Babokhov. “Dynamic chromatin organization without the 30-nm fiber”. In: *Current opinion in cell biology* 58 (2019), pp. 95–104.
- [16] Jill E Moore et al. “Expanded encyclopaedias of DNA elements in the human and mouse genomes”. In: *Nature* 583.7818 (2020), pp. 699–710.
- [17] Illumina. “Illumina Sequencing Technology”. In: (2010). URL: https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf.
- [18] Illumina. “An introduction to Next-Generation Sequencing Technology”. In: (2017). URL: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf.
- [19] Yan-Jie Feng et al. “Parallel tagged amplicon sequencing of relatively long PCR products using the Illumina HiSeq platform and transcriptome assembly”. In: *Molecular Ecology Resources* 16.1 (2016), pp. 91–102.
- [20] Alexander Payne et al. “BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files”. In: *Bioinformatics* 35.13 (2019), pp. 2193–2198.
- [21] Jason D Buenrostro et al. “Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position”. In: *Nature methods* 10.12 (2013), pp. 1213–1218.
- [22] William S Reznikoff. “Transposon tn 5”. In: *Annual review of genetics* 42 (2008), pp. 269–286.
- [23] Andrew Adey et al. “Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition”. In: *Genome biology* 11 (2010), pp. 1–17.
- [24] Maria Tsompana and Michael J Buck. “Chromatin accessibility: a window into the genome”. In: *Epigenetics & chromatin* 7.1 (2014), pp. 1–16.
- [25] Fiorella C Grandi et al. “Chromatin accessibility profiling by ATAC-seq”. In: *Nature Protocols* 17.6 (2022), pp. 1518–1552.
- [26] Haley M Amemiya, Anshul Kundaje, and Alan P Boyle. “The ENCODE blacklist: identification of problematic regions of the genome”. In: *Scientific reports* 9.1 (2019), p. 9354.
- [27] Darren A Cusanovich et al. “Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing”. In: *Science* 348.6237 (2015), pp. 910–914.
- [28] Jason D Buenrostro et al. “Single-cell chromatin accessibility reveals principles of regulatory variation”. In: *Nature* 523.7561 (2015), pp. 486–490.

- [29] Ansuman T Satpathy et al. “Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion”. In: *Nature biotechnology* 37.8 (2019), pp. 925–936.
- [30] M Ryan Corces et al. “Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution”. eng. In: *Nature genetics* 48.10 (2016), pp. 1193–1203.
- [31] Elisa Laurenti and Berthold Göttgens. “From haematopoietic stem cells to complex differentiation landscapes”. In: *Nature* 553.7689 (2018), pp. 418–426.
- [32] Seungbyn Baek and Insuk Lee. “Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation”. In: *Computational and structural biotechnology journal* 18 (2020), pp. 1429–1439.
- [33] Liesbeth Minnoye et al. “Chromatin accessibility profiling methods”. In: *Nature Reviews Methods Primers* 1.1 (2021), p. 10.
- [34] Huidong Chen et al. “Assessment of computational methods for the analysis of single-cell ATAC-seq data”. In: *Genome biology* 20.1 (2019), pp. 1–25.
- [35] Rongxin Fang et al. “Comprehensive analysis of single cell ATAC-seq data with SnapATAC”. In: *Nature communications* 12.1 (2021), p. 1337.
- [36] Jeffrey M Granja et al. “ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis”. In: *Nature genetics* 53.3 (2021), pp. 403–411.
- [37] Darren A Cusanovich et al. “The cis-regulatory dynamics of embryonic development at single-cell resolution”. In: *Nature* 555.7697 (2018), pp. 538–542.
- [38] Timothy J Durham et al. “Comprehensive characterization of tissue-specific chromatin accessibility in L2 *Caenorhabditis elegans* nematodes”. In: *Genome Research* 31.10 (2021), pp. 1952–1969.
- [39] Jason D. Buenrostro et al. “Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation”. In: *Cell* 173.6 (2018), 1535–1548.e16. DOI: <https://doi.org/10.1016/j.cell.2018.03.074>.
- [40] Darren A Cusanovich et al. “A single-cell atlas of in vivo mammalian chromatin accessibility”. In: *Cell* 174.5 (2018), pp. 1309–1324.
- [41] Carmen Bravo González-Blas et al. “cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data”. In: *Nature methods* 16.5 (2019), pp. 397–400.
- [42] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.
- [43] Matthew Hoffman, Francis Bach, and David Blei. “Online learning for latent dirichlet allocation”. In: *advances in neural information processing systems* 23 (2010).
- [44] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).
- [45] Han Yuan and David R Kelley. “scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks”. In: *Nature Methods* 19.9 (2022), pp. 1088–1096.

- [46] Alicia N Schep et al. “chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data”. In: *Nature methods* 14.10 (2017), pp. 975–978.
- [47] David R Kelley, Jasper Snoek, and John L Rinn. “Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks”. In: *Genome research* 26.7 (2016), pp. 990–999.
- [48] Morten Beck Rye, Pål Sætrom, and Finn Drabløs. “A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs”. In: *Nucleic acids research* 39.4 (2011), e25–e25.
- [49] Jianxing Feng et al. “Identifying ChIP-seq enrichment using MACS”. In: *Nature protocols* 7.9 (2012), pp. 1728–1740.
- [50] Naim U Rashid et al. “ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions”. In: *Genome biology* 12 (2011), pp. 1–20.
- [51] Alan P Boyle et al. “F-Seq: a feature density estimator for high-throughput sequence tags”. In: *Bioinformatics* 24.21 (2008), pp. 2537–2538.
- [52] John M. Gaspar. *Genrich: detecting sites of genomic enrichment*. URL: <https://github.com/jsh58/Genrich>.
- [53] Evan D Tarbell and Tao Liu. “HMMRATAC: a Hidden Markov Modeler for ATAC-seq”. In: *Nucleic acids research* 47.16 (2019), e91–e91.
- [54] Ha TH Vu et al. “Unsupervised contrastive peak caller for ATAC-seq”. In: *Genome Research* (2023), gr–277677.
- [55] Lance D Hentges et al. “LanceOtron: a deep learning peak caller for genome sequencing experiments”. In: *Bioinformatics* 38.18 (2022), pp. 4255–4263.
- [56] Gökçen Eraslan et al. “Deep learning: new computational modelling techniques for genomics”. In: *Nature Reviews Genetics* 20.7 (2019), pp. 389–403.
- [57] Timothy L Bailey et al. “MEME SUITE: tools for motif discovery and searching”. In: *Nucleic acids research* 37.suppl_2 (2009), W202–W208.
- [58] Tao Liu et al. “Cistrome: an integrative platform for transcriptional regulation studies”. In: *Genome biology* 12.8 (2011), pp. 1–10.
- [59] Ivan V Kulakovskiy et al. “HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models”. In: *Nucleic acids research* 44.D1 (2016), pp. D116–D125.
- [60] Jaime A Castro-Mondragon et al. “JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles”. In: *Nucleic acids research* 50.D1 (2022), pp. D165–D173.
- [61] Jason Ernst and Manolis Kellis. “Discovery and characterization of chromatin states for systematic annotation of the human genome”. In: *Nature biotechnology* 28.8 (2010), pp. 817–825.
- [62] Jason Ernst and Manolis Kellis. “ChromHMM: automating chromatin-state discovery and characterization”. eng. In: *Nature methods* 9.3 (2012), pp. 215–216.

- [63] Jason Ernst and Manolis Kellis. “Chromatin-state discovery and genome annotation with ChromHMM”. eng. In: *Nature protocols* 12.12 (2017), pp. 2478–2492.
- [64] Ha Vu and Jason Ernst. “Universal annotation of the human genome through integration of over a thousand epigenomic datasets”. In: *Genome biology* 23 (2022), pp. 1–37.
- [65] Michael M Hoffman et al. “Unsupervised pattern discovery in human chromatin structure through genomic segmentation”. In: *Nature methods* 9.5 (2012), pp. 473–476.
- [66] Kyung-Ah Sohn et al. “hiHMM: Bayesian non-parametric joint inference of chromatin state maps”. In: *Bioinformatics* 31.13 (2015), pp. 2066–2074.
- [67] David Van Dijk et al. “Recovering gene interactions from single-cell data using data diffusion”. In: *Cell* 174.3 (2018), pp. 716–729.
- [68] Mo Huang et al. “SAVER: gene expression recovery for single-cell RNA sequencing”. In: *Nature methods* 15.7 (2018), pp. 539–542.
- [69] Kexin Huang. “scGNN: scRNA-seq Dropout Imputation via Induced Hierarchical Cell Similarity Graph”. In: *arXiv preprint arXiv:2008.03322* (2020).
- [70] Juexin Wang et al. “scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses”. In: *Nature communications* 12.1 (2021), p. 1882.
- [71] Babak Alipanahi et al. “Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning”. In: *Nature biotechnology* 33.8 (2015), pp. 831–838.
- [72] Jian Zhou and Olga G Troyanskaya. “Predicting effects of noncoding variants with deep learning-based sequence model”. In: *Nature methods* 12.10 (2015), pp. 931–934.
- [73] David R Kelley et al. “Sequential regulatory activity prediction across chromosomes with convolutional neural networks”. In: *Genome research* 28.5 (2018), pp. 739–750.
- [74] Žiga Avsec et al. “Effective gene expression prediction from sequence by integrating long-range interactions”. In: *Nature methods* 18.10 (2021), pp. 1196–1203.
- [75] Praveen Weeratunga et al. “Unbiased single cell spatial analysis localises inflammatory clusters of immature neutrophils-CD8 T cells to alveolar progenitor cells in fatal COVID-19 lungs”. In: *MedRxiv* (2022), pp. 2022–12.
- [76] Martin J Sergeant et al. “Multi Locus View: an extensible web-based tool for the analysis of genomic data.” In: *Communications biology* 4.1 (2021), p. 623.
- [77] Johannes Köster and Sven Rahmann. “Snakemake—a scalable bioinformatics workflow engine”. In: *Bioinformatics* 28.19 (2012), pp. 2520–2522.
- [78] Blagoje Soskic et al. “Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases”. In: *Nature genetics* 51.10 (2019), pp. 1486–1493.
- [79] Damien J Downes et al. “An integrated platform to systematically identify causal variants and genes for polygenic human traits”. In: *BioRxiv* (2019), p. 813618.

- [80] William J Astle et al. “The allelic landscape of human blood cell trait variation and links to common complex disease”. In: *Cell* 167.5 (2016), pp. 1415–1429.
- [81] Avantika Lal et al. “Deep learning-based enhancement of epigenomics data with AtacWorks”. eng. In: *Nature communications* 12.1 (2021), pp. 1507–1507.
- [82] Illumina. “Nextera® Library Validation and Cluster Density Optimization”. In: (2014). URL: https://www.illumina.com/documents/products/technotes/technote_nextera_library_validation.pdf.
- [83] Yuichiro Hara et al. “Optimizing and benchmarking de novo transcriptome sequencing: from library preparation to assembly evaluation”. In: *BMC genomics* 16.1 (2015), pp. 1–12.
- [84] Teodora Ribarska et al. “Optimization of enzymatic fragmentation is crucial to maximize genome coverage: a comparison of library preparation methods for Illumina sequencing”. In: *BMC genomics* 23.1 (2022), p. 92.
- [85] Andrew Viterbi. “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm”. In: *IEEE transactions on Information Theory* 13.2 (1967), pp. 260–269.
- [86] Sven Heinz et al. “Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities”. In: *Molecular cell* 38.4 (2010), pp. 576–589.
- [87] Hideyuki Yoshida et al. “The cis-regulatory atlas of the mouse immune system”. In: *Cell* 176.4 (2019), pp. 897–912.
- [88] Steven Henikoff et al. “Efficient chromatin accessibility mapping in situ by nucleosome-tethered tagmentation”. In: *Elife* 9 (2020), e63274.
- [89] Gregory M Cooper et al. “Distribution and intensity of constraint in mammalian genomic sequence”. In: *Genome research* 15.7 (2005), pp. 901–913.
- [90] Adam Siepel, Katherine S Pollard, and David Haussler. “New methods for detecting lineage-specific selection”. In: *Research in Computational Molecular Biology: 10th Annual International Conference, RECOMB 2006, Venice, Italy, April 2-5, 2006. Proceedings 10*. Springer. 2006, pp. 190–205.
- [91] Joseph Felsenstein and Gary A Churchill. “A Hidden Markov Model approach to variation among sites in rate of evolution.” In: *Molecular biology and evolution* 13.1 (1996), pp. 93–104.
- [92] Adam Siepel et al. “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes”. In: *Genome research* 15.8 (2005), pp. 1034–1050.
- [93] Adam Siepel and David Haussler. “Computational identification of evolutionarily conserved exons”. In: *Proceedings of the eighth annual international conference on Research in computational molecular biology*. 2004, pp. 177–186.
- [94] JW Thomas et al. “Comparative analyses of multi-species sequences from targeted genomic regions”. In: *Nature* 424.6950 (2003), pp. 788–793.
- [95] Zhijian Li et al. “Identification of transcription factor binding sites using ATAC-seq”. In: *Genome biology* 20 (2019), pp. 1–21.

- [96] Mette Bentsen et al. “ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation”. In: *Nature communications* 11.1 (2020), p. 4267.
- [97] Anusri Pampari et al. *Bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants*. Version 0.1.1. Jan. 2023. DOI: 10.5281/zenodo.7567627. URL: <https://github.com/kundajelab/chrombpnet>.
- [98] Yan Hu et al. “Single-cell multi-scale footprinting reveals the modular organization of DNA regulatory elements”. In: *bioRxiv* (2023).
- [99] Bradley E Bernstein et al. “The NIH Roadmap Epigenomics Mapping Consortium”. eng. In: *Nature biotechnology* 28.10 (2010), pp. 1045–1048.
- [100] Angus M Sinclair. “Erythropoiesis stimulating agents: approaches to modulate activity”. In: *Biologics: Targets and Therapy* (2013), pp. 161–174.
- [101] Elisa Laurenti and Berthold Göttgens. “From haematopoietic stem cells to complex differentiation landscapes”. In: *Nature* 553.7689 (2018), pp. 418–426.
- [102] Julia Truch et al. “The chromatin remodeller ATRX facilitates diverse nuclear processes, in a stochastic manner, in both heterochromatin and euchromatin”. In: *Nature Communications* 13.1 (2022), p. 3485.
- [103] Wen Chen et al. “Abcb10 physically interacts with mitoferrin-1 (Slc25a37) to enhance its stability and function in the erythroid mitochondria”. In: *Proceedings of the National Academy of Sciences* 106.38 (2009), pp. 16263–16268.
- [104] Vivek Rai et al. “Single-cell ATAC-Seq in human pancreatic islets and deep learning upscaling of rare cells reveals cell-specific type 2 diabetes regulatory signatures”. In: *Molecular metabolism* 32 (2020), pp. 109–121.
- [105] Jeffrey M Granja et al. “Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia”. In: *Nature biotechnology* 37.12 (2019), pp. 1458–1465.
- [106] Tim Stuart et al. “Comprehensive integration of single-cell data”. In: *Cell* 177.7 (2019), pp. 1888–1902.
- [107] Yuhao Hao et al. “Integrated analysis of multimodal single-cell data”. In: *Cell* 184.13 (2021), pp. 3573–3587.
- [108] William M Rand. “Objective criteria for the evaluation of clustering methods”. In: *Journal of the American Statistical association* 66.336 (1971), pp. 846–850.
- [109] Nguyen Xuan Vinh, Julien Epps, and James Bailey. “Information theoretic measures for clusterings comparison: is a correction for chance necessary?” In: *Proceedings of the 26th annual international conference on machine learning*. 2009, pp. 1073–1080.
- [110] Andrew Rosenberg and Julia Hirschberg. “V-measure: A conditional entropy-based external cluster evaluation measure”. In: *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*. 2007, pp. 410–420.
- [111] Akshay Sanghi et al. “Chromatin accessibility associates with protein-RNA correlation in human cancer”. In: *Nature communications* 12.1 (2021), p. 5732.

- [112] Damien J. Downes et al. “High-resolution targeted 3C interrogation of cis-regulatory element organization at genome-wide scale”. eng. In: *Nature communications* 12.1 (2021), pp. 531–531.
- [113] Tim Stuart et al. “Single-cell chromatin state analysis with Signac”. In: *Nature methods* 18.11 (2021), pp. 1333–1341.
- [114] Lei Xiong et al. “SCALE method for single-cell ATAC-seq analysis via latent feature extraction”. In: *Nature communications* 10.1 (2019), p. 4576.
- [115] Hao Li et al. “Inferring transcription factor regulatory networks from single-cell ATAC-seq data based on graph neural networks”. eng. In: *Nature machine intelligence* 4.4 (2022), pp. 389–400.
- [116] Chenfei Wang et al. “Integrative analyses of single-cell transcriptome and regulome using MAESTRO”. In: *Genome biology* 21.1 (2020), pp. 1–28.
- [117] Wenbao Yu et al. “scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data”. In: *Genome biology* 21.1 (2020), pp. 1–17.
- [118] Florian V De Rop et al. “Systematic benchmarking of single-cell ATAC-sequencing protocols”. In: *Nature Biotechnology* (2023), pp. 1–11.
- [119] David Lähnemann et al. “Eleven grand challenges in single-cell data science”. In: *Genome biology* 21.1 (2020), pp. 1–35.