Tool Development and Computational Analysis of Chromatin with Applications to Renin Cell Biology

Jason Paul Smith

Charlottesville, Virginia

Masters of Science, Biological and Physical Sciences, University of Virginia, 2019

Masters of Science, Marine Biology, College of Charleston, 2014

Bachelors of Science, Biological Sciences, North Carolina State University, 2004

A Dissertation presented to the Graduate Faculty of the University of Virginia in Candidacy for the Degree of Doctor of Philosophy

Department of Biochemistry and Molecular Genetics

May, 2022

This work is dedicated to all the researchers and mentors who provided encouragement, criticism, and support along the way: Kevin Shianna PhD, Erin Richard PhD, Todd Stukenberg PhD, Aakrosh Ratan PhD, Shayn Peirce-Cottler PhD, Ariel Gomez PhD, and Nathan Sheffield PhD.

Most importantly, this work is in thanks to and in dedication to the person who made it all possible, my wife Amy Smith and our support team of Henry and Jack Smith.

List of Abbreviations

ACE Angiotensin Converting Enzyme
Acta2 Actin alpha 2, smooth muscle
Akr1b7 Aldo-keto reductase family 1 member B7
ATAC-seq Assay for Transposase Accessible Chromatin using high-throughput sequencing
Bach1 BTB Domain And CNC Homolog 1
Bach2 BTB domain and CNC homolog 2
bp Base pair
Bcl11a BAF Chromatin Remodeling Complex Subunit BCL11A
Bcl11b BAF Chromatin Remodeling Complex Subunit BCL11B
BSA Bovine Serum Albumin
cAMP Cyclic adenosine monophosphate
CBP CREB binding protein
cDNA Complementary deoxyribonucleic acid
ChIP-seq Chromatin immunoprecipitation assay with high-throughput sequencing
ChRO-seq Chromatin run-on sequencing assay
CM cap mesenchyme
Cnn1 Calponin 1
CRE Cyclic AMP response element
CREB cAMP response element-binding protein
Crip1 Cysteine rich protein 1
CTCF CCCTC-binding factor
dPBS Dulbecco's phosphate-buffered saline
DHS DNase I hypersensitive site
DMEM Dulbecco's Modified Eagle Medium
DNA Deoxyribonucleic Acid
DNase Deoxyribonuclease
DNase-seq DNase I hypersensitive sites sequencing
Dusp1 Dual Specificity Phosphatase 1
E12 Mouse embryonic day 12
E18 Mouse embryonic day 18
Ebf1 EBF transcription factor 1
EDTA Ethylenediaminetetraacetic acid
Ets1 ETS Proto-Oncogene 1, Transcription Factor
FAIRE-seq Formaldehyde-Assisted Isolation of Regulatory Elements using high-throughput sequencing
FBS Fetal bovine serum
FIMO Find Individual Motif Occurerences
Fos FOS proto-cncogene, AP-1 transcription factor subunit Fosl2 FOS Like 2, AP-1 Transcription Factor Subunit
Foxsl Forkhead Box S1
Foxd1 Forkhead Box D1
GEM Cell-Gel Beads in Emulsion
GFP Green fluorescent protein
GRO-seq Global run-on sequencing assay
GTF General transcription factor
H2A Histone 2A
H2B Histone 2B
H3 Histone 3
H4 Histone 4
H3K4me1 Histone 3 Lysine 4 monomethylation
H3K4me3 Histone 3 Lysine 4 trimethylation
H3K27ac Histone 3 Lysine 27 acetylation
HEPES 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
JG juxtaglomerular
Jun Jun Proto-Oncogene, AP-1 Transcription Factor Subunit
LSI Latent Semantic indexing

MARE Maf-recognition elements
MEF2 Myocyte enhancer factor 2 family of transcription factors
Mef2a Myocyte enhancer factor 2a
Mef2b Myocyte enhancer factor 2b
Mef2c Myocyte enhancer factor 2c
Mef2d Myocyte enhancer factor 2d
miRNA Micro ribonucleic acids
MNase Micrococcal nuclease
MNase-seq Micrococcal nuclease digestion with deep sequencing
mRNA Messenger ribonucleic acids
mt Mitochondria
Myh11 Myosin, heavy polypeptide 11, smooth muscle
Nfic Nuclear factor I C
Nfix Nuclear factor I X
Nfya Nuclear transcription factor Y subunit alpha
NLP Natural language processing
Nr2c1 Nuclear receptor subfamily 2 group c member 1
Nr2f6 Nuclear receptor subfamily 2 group f member 6
p300 Histone acetyltransferase p300
P5 Mouse postnatal day 5
P30 Mouse postnatal day 30
PC pericyte
PCA Principal component analysis
PCR Polymerase chain reaction
PEP Portable Encapsulated Project
PEPATAC Portable Encapsulated Project for ATAC-seq analysis
PEPPRO Portable Encapsulated Project for PRO-seq analysis
PFM Position frequency matrices
PRO-seq Precision run-on sequencing assay
PCT proximal convoluted tubule
PT proximal tubule
Rarg Retinoic acid receptor gamma
RAS Renin-Angiotensin System
RBP-J Recombination signal binding protein for immunoglobulin kappa J region
rCRSd revised Cambridge Reference Sequence doubled genome
RD Rapidly dividing
RE regulatory element
Ren1 Renin-1
Rfx2 Regulatory factor X2
RNA Ribonucleic acid
RNA pol II Ribonucleic acid polymerase II
RNA-seq acid high-throughput sequencing
rRNA Ribosomal ribonucleic acid
rDNA Ribosomal deoxyribonucleic acid
scATAC-seq Single cell Assay for Transposase Accessible Chromatin using high-throughput sequencing
scRNA-seq Single cell assay for ribonucleic acid using high-throughput sequencing
SE Super enhancer
Smarcc1 SWI/SNF related, matrix associated, actin dependent regulator of chromatin subfamily C member 1
SMC Smooth muscle cell
Smtn Smoothelin
Stat3 Signal Transducer And Activator Of Transcription 3
Stat5b Signal Transducer And Activator Of Transcription 5B
Tagln Transgelin
Tcf/Lef T cell factor/lymphoid enhancer factor family
Tead3 TEA Domain Transcription Factor 3
TGF-β1 Transforming growth factor beta 1

TF Transcription factor
tRNA Transfer ribonucleic acids
TSS Transcription start site
UMAP Uniform manifold approximation and projection
UMI Unique molecular identifier
VEC vascular endothelial cell
VSMC vascular smooth muscle cell
Zfp283 Zinc finger protein 283
Zfp36 Zinc finger protein 36
Zfp384 Zinc finger protein 384

Tool Development and Computational Analysis of Chromatin with Applications to Renin Cell Biology

1 Introduction **1**
- 1.1 Chromatin . . . 1
  - 1.1.1 Chromatin overview . . . 1
  - 1.1.2 Chromatin structure . . . 1
    - 1.1.2.1 Nucleosomes . . . 2
    - 1.1.2.2 Histones . . . 2
  - 1.1.3 Regulatory regions . . . 3
    - 1.1.3.1 Promoters . . . 3
    - 1.1.3.2 Enhancers . . . 4
    - 1.1.3.3 Insulators . . . 5
    - 1.1.3.4 Silencers . . . 5
  - 1.1.4 Transcription factors . . . 5
- 1.2 Assays for measuring open chromatin . . . 5
  - 1.2.1 MNase-seq . . . 6
  - 1.2.2 DNase-seq . . . 6
  - 1.2.3 FAIRE-seq . . . 6
  - 1.2.4 ATAC-seq . . . 7
  - 1.2.5 scATAC-seq . . . 7
- 1.3 Assays for measuring gene expression . . . 8
  - 1.3.1 RNA-seq . . . 8
  - 1.3.2 scRNA-seq . . . 8
  - 1.3.3 Nascent RNA-seq . . . 8
- 1.4 Computational challenges of genomic and epigenomic analysis . . . 9
  - 1.4.1 Bioinformatic analysis of single-cell genomics . . . 9
- 1.5 Renin cell development . . . 9
  - 1.5.1 Renin cell overview . . . 10
  - 1.5.2 Current knowledge of renin cell regulation . . . 11

2 Analytical approaches for ATAC-seq data analysis (modified from [213]) **12**
- 2.1 Introduction . . . 12
- 2.2 Fundamentals of ATAC-seq Data Analysis . . . 14
- 2.3 Alignment, adapters, and mitochondrial reads . . . 14
- 2.4 Removing duplicates . . . 15
- 2.5 Generating signal tracks . . . 16
- 2.6 Peak calling . . . 16
- 2.7 Downstream analysis . . . 16
- 2.8 Survey of Tools for ATAC-seq Analysis . . . 17
  - 2.8.1 Step-by-step Analysis Guides . . . 17
  - 2.8.2 Raw Sequence Pipelines and Workflows . . . 18
  - 2.8.3 Quality control . . . 20
  - 2.8.4 Peak calling . . . 20
  - 2.8.5 Differential accessibility . . . 21
  - 2.8.6 Motif enrichment and TF footprinting . . . 22
  - 2.8.7 Nucleosome positioning . . . 23
  - 2.8.8 Region enrichment . . . 25
  - 2.8.9 Single-cell . . . 25
- 2.9 Conclusion . . . 27

3 PEPPRO: quality control and processing of nascent RNA profiling data (modified from [215]) **27**
- 3.1 Background . . . 27
- 3.2 Results . . . 28
  - 3.2.1 Pipeline overview and data description . . . 28
  - 3.2.2 Adapter ratio . . . 30
  - 3.2.3 RNA integrity . . . 31

3.2.4 Library complexity . . . 32
3.2.5 Nascent RNA purity . . . 33
3.2.6 Run-on efficiency . . . 34
3.2.7 Read feature distributions . . . 36
3.2.8 Differential expression . . . 36
3.2.9 Metric robustness . . . 37
3.3 Conclusions . . . 38
3.4 Availability of data and materials . . . 38
3.5 Methods . . . 39
3.5.1 Pipeline implementation . . . 39
3.5.2 Refgenie reference assembly resources . . . 39
3.5.3 Adapter-adapter ligation product abundance . . . 40
3.5.4 RNA insert size distribution and degradation . . . 40
3.5.5 Excluding size selection skews metrics . . . 41
3.5.6 Removing UMI and reverse complementation . . . 41
3.5.7 Serial alignments . . . 42
3.5.8 Processed signal tracks . . . 42
3.5.9 Exon-intron ratio plots . . . 43
3.5.10 Pause index . . . 43
3.5.11 PRO-seq experiments . . . 44
3.5.12 Synthetic experiments . . . 44
3.6 Supplemental . . . 45
3.6.1 R code to generate a gene counts table . . . 45
3.6.2 Supplemental figures . . . 45

4 PEPATAC: An optimized pipeline for ATAC-seq data analysis with serial alignments (modified from [177]) 55
4.1 Introduction . . . 55
4.2 Materials and Methods . . . 56
4.2.1 PEPATAC configuration . . . 56
4.2.2 Refgenie reference assembly resources . . . 58
4.2.3 File inputs and adapter trimming . . . 60
4.2.4 Prealignments and mitochondrial DNA . . . 60
4.2.5 Alignments, deduplication, and library complexity . . . 61
4.2.6 Library QC metrics . . . 61
4.2.7 Signal tracks and peak calling . . . 62
4.2.8 Running multiple samples with `PEPATAC` . . . 63
4.2.9 Aggregating results from multiple samples . . . 63
4.3 Results . . . 64
4.3.1 Performance . . . 64
4.3.2 Prealignments . . . 64
4.3.3 Peak caller comparison . . . 67
4.3.4 Library QC comparison . . . 68
4.3.5 Fraction of reads in peaks . . . 68
4.4 Discussion . . . 69
4.4.1 Documentation and links . . . 69
4.5 Supplemental . . . 69
4.5.1 Supplemental figures . . . 69
4.5.2 Supplemental_file_1 . . . 75
4.5.3 Supplemental_file_2 . . . 75
4.5.4 Supplemental_file_3 . . . 75
4.5.5 Supplemental_file_4 . . . 75

5 MEF2 family of transcription factors contribute to renin cell identity 76
5.1 Background . . . 76
5.2 Results . . . 77
5.2.1 Overview of the epigenetic landscape of juxtaglomerular cell development 78
5.2.2 Differentiation trajectory of juxtaglomerular cells . . . 80

5.2.3 Transcription factors contributing to juxtaglomerular cell development 82
5.2.4 MEF2 family of TFs separates JG cells from mature SMCs 84
5.3 Discussion 86
5.4 Materials and Methods 87
5.4.1 Mouse model (from the UVA Pediatric Center of Excellence in Nephrology) 87
5.4.2 Isolation of kidney cells 87
5.4.2.1 Isolation of kidney single cells: E12 88
5.4.2.2 Isolation of kidney single cells: E18, P5 and P30 88
5.4.3 scATAC-seq library preparation 89
5.4.4 scRNA-seq library preparation 89
5.4.5 Read preprocessing 90
5.4.5.1 Genome and transcriptome annotations 90
5.4.5.2 scATAC-seq alignment and fragment matrix generation . . . 90
5.4.5.3 scATAC-seq quality control 90
5.4.5.4 scATAC-seq dimensionality reduction and batch correction . 91
5.4.5.5 scATAC-seq cell clustering 91
5.4.5.6 scRNA-seq alignment and feature-barcode matrix generation 91
5.4.5.7 scRNA-seq quality control 91
5.4.5.8 scRNA-seq cell clustering 92
5.4.5.9 scRNA-seq cell identification 92
5.4.6 Integrating transciptome and accessibilome 92
5.4.7 Renin cell differentiation trajectory 93
5.4.7.1 Identifying renin cells 93
5.4.8 Pairwise comparisons of renin trajectory cells 94
5.4.8.1 Renin trajectory marker peaks 94
5.4.8.2 Motif annotations and enrichment 94
5.4.8.3 TF footprinting 94
5.4.8.4 Peak co-accessibility and peak to gene links 95
5.4.8.5 Positive transcription factor regulators 95
5.5 Supplemental figures 96

6 Companion research that improves our ability to interrogate genomic regions 104
6.1 Looper: A pipeline submission engine 105
6.2 Refgenie: a reference genome resource manager (derived from [448]) 105
6.3 Embeddings of genomic region sets capture rich biological associations in lower dimensions (derived from [454]) 105
6.4 GenomicDistributions: fast analysis of genomic intervals with bioconductor (derived from [458]) 106

7 Conclusions, Impact, and Future Studies 107
7.1 Summary of fulfilled gaps in the field 107
7.2 Integration of computational and bench-based methods 108
7.3 Support for analytical tools 108
7.4 Investigating the development and regulation of renin cells 108
7.4.1 Epigenetic regulation of early, middle, and late progenitors of renin cells 109
7.4.2 Identifying druggable targets of aberrant renin cell function and renin expression 109

8 Materials and Methods 109
8.1 Computational infrastructure 110
8.2 Building a better pipeline 110

9 References 112

List of Figures

Figure 1.1 **Histone modifications are indicative of promoter and enhancer activity**. 3

Figure 1.2 **Distribution of renin cells in development and the role of renin in homeostatic control**. (a) Renin cells line afferent arterioles during embryogenesis before differentiating into vascular smooth muscle cells (VSMC) and mesangial cells with mature renin-expressing cells restricted to the juxtaglomerular region in adult mammals. (b) Renin converts angiotensinogen into angiotensin I to initiate homeostatic balance. . . . 10

Figure 1.3 **Super-enhancers (SEs) act as chromatin sensors that control the identity and memory of renin cells to maintain homeostasis.** Schematic summarizing the main signaling pathways, and chromatin changes involved in the maintenance of juxtaglomerular (JG) cell identity and reacquisition of the renin phenotype by smooth muscle cells (SMCs) in response to physiological demands. Activation of the cAMP or the Notch pathways leads to profound epigenetic changes at the renin locus regulatory region characterized by deposition of acetylation of lysine 27 of H3 by p300, sliding of nucleosomes, and opening of chromatin, which facilitate the access of numerous transcription factors including but not limited to Med1 (Mediator complex 1), Creb (cAMP-responsible element binding protein) 1, and RBP-J (recombination signal binding protein for immunoglobulin kappa J region). Loop formation is maintained by Ctcf (CCCTC-binding factor). The colored dots indicate the presence of additional SEs throughout the genome that also regulate renin cell identity. β-AR indicates beta adrenergic receptor; AC, adenylate cyclase; CREBP, phosphorylated cAMP-responsive element binding protein; EP4, Prostaglandin E2 receptor 4; Gsα, activating G-protein–coupled subunit; NICD, Notch intracellular domain; PGE2, prostaglandin E2; PKA, protein kinase A; Pol II, RNA polymerase II; and RBP-J, recombination signal binding protein for immunoglobulin kappa J region. Illustration credit: Ben Smith. Data derived from Martinez et al. [228]. Included by permission from Maria Luisa S. Sequeira-Lopez and R. Ariel Gomez. Renin Cells, the Kidney, and Hypertension. Circulation Research. Volume: 128, Issue: 7, Pages: 887-907, DOI:10.1161/CIRCRESAHA.121.318064. and the publisher, Wolters Kluwer Health, Inc. Please contact permissions@lww.com for further information. 11

Figure 2.1 **ATAC-seq is a rapidly growing method for open chromatin analysis.** (a) Increasing prevalence of 'ATAC-seq' DataSets in the Gene Expression Omnibus (GEO) (Color = Species, Gray line = fitted exponential growth model) (b) Generalized ATAC-seq library prep protocol 13

Figure 2.2 **ATAC-seq general workflow.** Raw reads are processed through a series of steps to produce uniform intermediate results, which can then be further analyzed with more analysis specific to a biological research question. 15

Figure 3.1 **`PEPPRO` test set data table and signal tracks**. A) Table showing the attributes of samples collected for our test set. Complete metadata is available from the `PEPPRO` website. B) Read count normalized signal tracks from published data are visualized within a browser (Scale is per 1M). 28

Figure 3.2 **`PEPPRO` steps for genomic run-on data.** `PEPPRO` starts from raw sequencing reads and produces a variety of quality control plots and processed output files for more detailed downstream analysis. 29

Figure 3.3 **RNA integrity is assessed with degradation ratios and insert sizes.** A) Schematic illustrating intact versus degraded libraries. B) Degradation ratio for test samples (HelaS3 GRO sample could not be calculated; Values less than dashed line (1.0) are considered high quality). C-F) Insert size distributions for: C, a degraded single-end library; D, a degraded paired-end library; E, a non-degraded single-end library; and F, a non-degraded paired-end library (orange shading represents highly degraded reads; yellow shading represents partially degraded reads). . . 30

Figure 3.4 **Library complexity is measured with unique read frequency distributions and projections.** A) Schematic demonstrating PCR duplication and library complexity (dashed line represents completely unique library). B) Library complexity traces plot the read count versus externally calculated deduplicated read counts. Deduplication is a prerequisite, so these plots may only be produced for samples with UMIs. Inset zooms to region from 0 to double the maximum number of unique reads. C) The position of curves in panel B at a sequencing depth of 10 million reads (dashed line represents minimum recommended percentage of unique reads). 31
Figure 3.5 **Nascent RNA purity is assessed with the exon-intron ratio**. A) Schematic demonstrating mRNA contamination calculation. X represents the exclusion of the first exon in the calculation. B) Median mRNA contamination metric for test set samples (Shaded region represents recommended range (1-1.8)). C) Histogram showing the distribution of mRNA contamination score across genes in the K562 PRO-seq sample. D) As in panel C for a GRO-seq library. E) mRNA contamination distribution for K562 PRO-seq spiked with 30% K562 RNA-seq. F) mRNA contamination distribution for HelaS3 GRO-seq is comparable to the 30% RNA-seq spike-in sample. $\tilde{x}$ = median; $\bar{x}$= mean) 32
Figure 3.6 **Run-on efficiency is measured with pause indices**. A) Schematic demonstrating pause index calculation. B) Pause index values for *Drosophila melanogaster* GRO-seq libraries with (GSM577247) or without sarkosyl (GSM577248). C) The histogram of pause index values is shifted to the right upon addition of sarkosyl in GRO-seq libraries. D) Pause index values for test set samples (Values above the dashed line are recommended). E) High pause index identified in H9 treated PRO-seq. F) Low pause index from HelaS3 GRO-seq. ($\tilde{x}$ = median; $\bar{x}$=mean) . . 33
Figure 3.7 **Fraction of reads in genomic features**. A) K562 PRO-seq represents a "good" cumulative fraction of reads in features (cFRiF) and fraction of reads in features (FRiF) plot. B) K562 PRO-seq with 90% K562 RNA-seq spike-in represents a "bad" FRiF/PRiF. 35
Figure 3.8 **Differential analysis with the `PEPPRO` counts matrix**. A) MA plot between H9 DMSO versus H9 200nM romidepsin treated PRO-seq libaries (dots = genes; top 10 most significant genes labeled; n=3/treatment). B) Most significantly differential gene count differences. C) Read count normalized signal tracks from the differential analysis (Scale is per 1M). 35
Figure 3.9 **Recommendation table.** Based on our experience processing both high- and low-quality nascent RNA libraries, these are our recommended values for high-quality PRO-seq libraries. 37
Figure 3.10 **K562 RNA-seq spike-in signal tracks show increasing exonic coverage**. GAPDH exonic coverage is enriched as the percentage of RNA-seq reads increases, and is visualized particularly well at exons 6 and 8. Each sample library is composed of 70M total reads. Scale for each track is 1000 to -20. 46
Figure 3.11 **Percentage of uninformative adapter reads following adapter removal for test set samples**. The HEK and H9 libraries contain more adapter-adapter reads because PAGE-mediated size selection was excluded from the protocol (Values below the dashed line are generally recommended for PAGE-purified libraries. Shaded region represents the recommended abundance of adapter reads for libraries *without* PAGE purification.) 46
Figure 3.12 **Ribosomal DNA alignment rates for test set samples**. The HelaS3 GRO-seq sample is highly enriched for ribosomal RNA transcripts compared to other test samples. Values below the dashed line are recommended. 47

Figure 3.13 **Abundance of rDNA to total reads is correlated with mature RNA contamination.** Correlation plot between the measure of mRNA contamination (median exon:intron density) and the ratio of rDNA aligned reads to total reads for: A) all primary samples (*excludes RNA-seq spike-in experiment due to ribosomal depletion inherent in RNA-seq library preparation), B) primary samples excluding the known outlier HelaS3_GRO-seq sample, C) all samples in panel B *and* all non-redundant samples from GSE126919 including three cellular subclones (A, B, and C) to demonstrate possible differences due to cell lines. Test for association determined with Pearson's product moment correlation coefficient. 47
Figure 3.14 **TSS enrichment**. A) TSS enrichment scores for test set samples. B) Representative high-quality PRO-seq TSS enrichment plot. C) TSS enrichment plot in romidepsin treated PRO-seq library. D) Representative high quality GRO-seq TSS enrichment plot E) Representative example of lower quality GRO-seq TSS enrichment plot. 48
Figure 3.15 **Fraction of Reads in Features in ChRO-seq**. Cumulative FRiF and FRiF (inset) plots for example Jurkat ChRO-seq 1 library test sample shows increased enrichment of promoter sequences. 48
Figure 3.16 **RNA-seq spike-in shows how FRiF changes with mRNA contamination**. A) Increasing percentages of RNA-seq spike-in lead to changes in the fraction of reads in features (FRiF). B) Cumulative FRiF plots at 10%, 50%, and 90% RNA-seq spike-in. Plot insets represent the expected versus observed fraction of reads in genomic features. Each spike-in library contains 70 million total reads. 49
Figure 3.17 **K562 PRO-seq signal tracks show increasing coverage with depth**. Incrementally subsampled K562 PRO-seq library signal tracks display reduced relative coverage at a representative locus (GAPDH). Fixed scale for each track is 100 to -10. 50
Figure 3.18 **H9 PRO-seq 2 signal tracks show increasing coverage with depth**. Incrementally subsampled H9 PRO-seq 2 library signal tracks display reduced relative coverage at a representative locus (GAPDH). Fixed scale for each track is 10 to -5. 50
Figure 3.19 **QC metrics are not affected by sequencing depth in subsampled K562 PRO-seq**. Using subsampled K562 PRO-seq data, we show how various metrics behave across a spectrum of sequencing depths: A) Degradation ratio, B) mRNA contamination, C) Pause index, D) the percentage of uninformative adapter reads, E) the rDNA alignment rate, F) and the TSS enrichment scores are unaffected by sequencing depth. G) The FRiF and cumulative FRiF is unaffected by sequencing depth. The complete K562 PRO-seq library (100%) contains approximately 497 million reads. 51
Figure 3.20 **QC metrics are not affected by sequencing depth in subsampled H9 PRO-seq**. Using subsampled H9 PRO-seq data, we show how various metrics behave across a spectrum of sequencing depths: A) Degradation ratio, B) mRNA contamination, C) Pause index, D) the percentage of uninformative adapter reads, E) the rDNA alignment rate, F) and the TSS enrichment scores are unaffected by sequencing depth. G) Library complexity traces plot the read count versus externally calculated deduplicated read counts. Deduplication is a prerequisite, so these plots may only be produced for samples with UMIs. Inset zooms to region from 0 to double the maximum number of unique reads. The position of curves in the left panel at a sequencing depth of 10 million reads (dashed line represents minimum recommended percentage of unique reads). H) The FRiF and cumulative FRiF is unaffected by sequencing depth. The complete H9 PRO-seq library (100%) contains approximately 116 million reads. 52

Figure 3.21 **QC metrics are not affected by low library complexity**. Using a synthetic set of libraries, we show how various metrics behave across a spectrum of complexity: A) Degradation ratio, B) mRNA contamination, C) Pause index, D) the percentage of uninformative adapter reads, E) the rDNA alignment rate, F) and the TSS enrichment scores are unaffected by low complexity. G) Library complexity traces plot the read count versus externally calculated deduplicated read counts. Deduplication is a prerequisite, so these plots may only be produced for samples with UMIs. Inset zooms to region from 0 to double the maximum number of unique reads. The right panel represents the position of curves in the left panel at a sequencing depth of 10 million reads (dashed line represents minimum recommended percentage of unique reads). *Libraries with less than 90% uniqueness could not be extrapolated due to saturation. H) The FRiF and cumulative FRiF is unaffected by low complexity. Each library contains 30 million total reads. 53

Figure 3.22 **Alternate annotation sources do not affect mRNA contamination and pause index**. A) The mRNA contamination metric and B) the pause index metric are robust across annotations. 54

Figure 4.1 **`PEPATAC` is feature-rich with a logical workflow**. (a) We compared features across 14 ATAC-seq pipelines (AIAP [342]; ATAC2GRN [343]; ATAC-pipe [344]; ATACProc [345]; CIPHER [346]; ENCODE [347]; esATAC [348]; GUAVA [349]; I-ATAC [350]; nfcore/atacseq [351]; pyflow-ATAC-seq [352]; seq2science [353]; snakePipes [354]; Tobias Rausch [355]) and `PEPATAC` stands out for being feature-rich . (b) Reads are preprocessed, serially aligned to the mitochondrial genome, curated repeats, and then the nuclear genome. `PEPATAC` generates both smooth and exact signal plots, called peaks, and QC output plots and tables. . 57

Figure 4.2 **Example `PEPATAC` QC plots for reads and peaks**. (a) Library complexity plots the read count versus externally calculated deduplicated read counts. Red line is library complexity curve for SRR5427743. Dashed line represents a completely unique library. Red diamond is the externally calculated duplicate read count. (b) TSS enrichment quality control plot. (c) Fragment length distribution showing characteristic peaks at mono-, di-, and tri-nucleosomes. (d) Cumulative fraction of reads in annotated genomic features (cFRiF). Inset: Fraction of reads in those features (FRiF). e) Signal tracks including: nucleotide-resolution and smoothed signal tracks. `PEPATAC` default peaks are called using the default pipeline settings for MACS2 [255]. (f) Distribution of peaks over the genome. (g) Distribution of peaks relative to TSS. (h) Distribution of peaks in annotated genomic partitions. Data from SRR5427743. . . 59

Figure 4.3 **`PEPATAC` prealignments increase mapped mtDNA reads, improve computational efficiency, and positively influences the fraction of reads in peaks (FRiP) metric.** (a) NuMTs represent a significant complication of simultaneous alignment. (b) At mtDNA percentages from 10-100% at total read numbers ranging from 10-200M, using prealignments dramatically reduces run time. (c) Log ratio of prealignments runtimes versus no prealignment runtimes yields significant savings. (d) There is a significant increase in the percent of reads mapped to mitochondrial sequence when using prealignments versus not across standard, fast, and omni-ATAC protocols. (e) As reported for ChIP-seq [371], FRiP is positively correlated with the number of called peaks. (f) With prealignments, the positive correlation between FRiP and the number of called peaks tends to increase ((d) ** = $p < 0.001$; t-test (mu = 0) with Benjamini-Hochberg correction. (e-f):* = $p < 0.0001$; Kendall rank correlation coefficient). 65

Figure 4.4 **ATAC-seq pipelines universally require several common bioinformatic tools.** While all pipelines require a number of common bioinformatic tools, PEPATAC offers the greatest flexibility and includes a number of the most popular tools. 70
Figure 4.5 **Deploying PEPATAC across multiple samples using looper**. The PEPATAC pipeline can be easily run across multiple samples in any computing environment using looper. 71
Figure 4.6 **PEPATAC is computational efficient**. (a) Pipeline runtime scales linearly with input file size. (b) Pipeline memory use peaks between 5-9GB. 71
Figure 4.7 **Prealignment increases mtDNA alignment.** Within Standard (a), Fast (b), and Omni (c) ATAC-seq library preparation protocols, every sample shows increased mtDNA alignment when utilizing prealignments (The gray lines represent the mean increase within each protocol. ** = $p < 0.001$; t-test (mu = 0) with Benjamini-Hochberg correction.) . . . 71
Figure 4.8 **Prealignment (and improved ATAC-seq library preparation protocols) successfully deplete signal from NuMTs, repeat regions, and high signal regions**. (a) Even where improved library preparation protocol leads to a NuMT annotated peak, prealignment successfully removes the spurious signal. (b) Both omni ATAC and prealignment to mitochondria *and* repeats and ribosomal sequence successfully depletes a spurious signal. 72
Figure 4.9 **Peaks are comparatively dissimilar between the five optional peak callers**. (a) For a single sample, MACS2 derived peaks, both with fixed and variable width peaks, are the most similar to Fseq called peaks. Genrich and HMMRATAC are the most unique among peak callers. (b) After PEPATAC consensus peak generation, HMMRATAC becomes even more dissimilar from the results derived from alternative peak callers. 73
Figure 4.10 **The TSS enrichment score is dependent on the annotation source**. Refgene TSS annotations, which include the predominant TSS annotation only, produces the highest TSS enrichment score. 73
Figure 4.11 **Prealignment changes the relationship between primary genome and total aligned reads and the fraction of reads in peaks (FRiP) is dependent on mapping strategy.** (a) The number of primary, nuclear genome mapped reads is reduced when using prealignments. (b) However, the total number of mapped reads is increased with prealignments due to more specific read mapping. (c) The FRiP is increased with prealignments when using primary, nuclear genome mapped reads as the denominator. (d) In contrast, when using the total mapped reads the FRiP is reduced when using prealignments due to a larger mapped read pool in the denominator (* = $p < 0.01$; ** = $p < 0.001$; t-test (mu = 0) with Benjamini-Hochberg correction). 74

Figure 5.1 **Overview of the experimental design to identify the renin cell developmental trajectory** . (a) Foxd1 progenitors in the cap mesenchyme in early E12 differentiate through E18, P5, and P30 to lead to mature renin expressing cells in the juxtaglomerular region. (b) Kidneys isolated from Foxd1-CRE recombinase mouse lineage-tracing model are sorted on Foxd1-derived GFP expression and single-cell ATAC-seq and RNA-seq is performed. (c) UMAP visualization of scATAC-seq data separated by time point. (d) UMAP visualization of scRNA-seq data with annotated cell clusters. (e) UMAP visualization of gene activity scores for canonical JG markers Ren1 and Akr1b7. (f) UMAP visualization of gene expression for canonical JG markers Ren1 and Akr1b7. (g) UMAP visualization of integrated scATAC-seq and scRNA-seq data with annotated cell clusters. (h) Cell frequency distribution across developmental time point. Numbers below timepoints represent total number of single cells at each time point. CM: cap mesenchyme; CD: collecting duct; JG: juxtaglomerular; PC: pericyte; EC: endothelial cell; SMC: smooth muscle cell; PT: proximal tubule; PCT: proximal convoluted tubule; RD: rapidly dividing. 79

Figure 5.2 **Epigenomic differentiation trajectory uncovers renin cells by post-natal day 5 in mouse kidney development**. (a) UMAP visualization showing the pseudo-time trajectory across developmental time points. Arrow head represents the end point of the trajectory. (b) Heatmap of motif hypergeometric enrichment-adjusted P values within the marker peaks of each JG trajectory cluster. Color indicates the motif enrichment (-log10(P value)) based on the hypergeometric test. (c) Integrated pseudo-time analysis of positively correlated gene scores and corresponding motifs. (d) Integrated pseudo-time analysis of positively correlated gene expression and corresponding motifs. Bold text indicates genes and motifs identified in both integrated approaches. CM: cap mesenchyme; CD: collecting duct; EC: endothelial cell; JG: juxtaglomerular; PC: pericyte; RD: rapidly dividing; SMC: smooth muscle cell; VEC: vascular endothelial cell; VSMC: vascular smooth muscle cell 81

Figure 5.3 **TF expression, accessibility, and enrichment uncover patterns of differentiation**. Footprints for Smarcc1 (a), Nfix (b), Mef2c (c), and Bach2 (d) are enriched in the JG cluster and parent clusters of JG and SMC cells. The expression pattern of enriched TFs in the JG cluster illustrates early (e), middle (f), late (g), and cyclical (h) patterns of transcript abundance. Gene score activity recapitulates expression patterns of early (e), middle (f), late (g), and cyclical (h) activity. TF deviation scores highlight enrichment of Smarcc1 (m), Nfix (n), Mef2c (o), and Bach2 (p) in late differentiation along the JG trajectory. CM: cap mesenchyme; CD: collecting duct; EC: endothelial cell; JG: juxtaglomerular; PC: pericyte; RD: rapidly dividing; SMC: smooth muscle cell; VEC: vascular endothelial cell; VSMC: vascular smooth muscle cell 83

Figure 5.4 **MEF2 family of TFs uniquely defines the terminal JG population**. (a). MA plot of the differential regulatory regions between SMCs and JG cells. (b) Top enriched motifs in SMCs as compared to JG cells. (c) Top enriched motifs in JG cells compared to SMCs. Mef2a (d), Mef2b (e), Mef2c (f), and Mef2d (g) footprints are enriched in JG cells. Mef2a expression (h) and gene score activity (l) peaks just prior to terminal differentiation into JG cells. Mef2b expression peaks in early differentiation (i) with the corresponding gene score activity steady early to middle before plummeting at differentiation into JG cells (m). Mef2c (j,n) and Mef2d (k,o) expression and gene activity both peak at terminal JG differentiation. (p) Browser tracks identify preferentially enriched open chromatin in the JG cell cluster at Ren1. Motif occurrences of enriched TFs identify putative binding sites in open and co-accessible peaks. Yellow fill box highlights uniquely enriched peak in JG cluster. Pink fill box highlights JG promoter region. JG: juxtaglomerular; SE: super-enhancer; SMC: smooth muscle cell 85

Figure 5.5 **scATAC-seq quality control**. (a-d). UMAP visualiztion of putative doublets in developmental time points E12 (a), E18 (b), P5 (c), and P30 (d). (e-h) Distributions of TSS enrichment by the log10 number of unique fragments for E12 (e), E18 (f), P5 (g), and P30 (h). Dashed lines represent cut off values below which cells are removed. (i-l) Fragment distribution plots for developmental time points E12 (i), E18 (j), P5 (k), and P30 (l). 96

Figure 5.6 **scRNA-seq quality control**. (a-d) Distribution of the number of RNA features against the total RNA count in developmental time points E12 (a), E18 (b), P5 (c), and P30 (d). (e-h) Distribution of the percentage of mitochondria mapped reads against the total RNA count in developmental time points E12 (e), E18 (f), P5 (g), and P30 (h). (i-l) Distribution of the percentage of hemoglobin mapped reads against the total RNA count in developmental time points E12 (i), E18 (j), P5 (k), and P30 (l). Green fill boxes represent cells passing filters. 97

Figure 5.7 **Browser tracks at the Ren1 locus identify previously reported promoter binding factor motif occurrences**. Browser tracks identify preferentially enriched open chromatin in the JG cell cluster at Ren1. Motif occurrences of enriched TFs identify putative binding sites in open and co-accessible peaks. Yellow fill box highlights uniquely enriched peak in JG cluster. Pink fill box highlights JG promoter region. 98

Figure 5.8 **Browser tracks at the Ren1 locus identify previously reported enhancer binding factor motif occurrences**. Browser tracks identify preferentially enriched open chromatin in the JG cell cluster at Ren1. Motif occurrences of enriched TFs identify putative binding sites in open and co-accessible peaks. Yellow fill box highlights uniquely enriched peak in JG cluster. Pink fill box highlights JG promoter region. 99

Figure 5.9 **Heatmaps show changes in differential accessibility, expression, and TF motif enrichment along the JG trajectory**. (a) Heatmap of gene score activity along cell clusters defining the JG trajectory. (b) Heatmap of gene expression along cell clusters defining the JG trajectory. (c) Heatmap of accessible regions identified along cell clusters defining the JG trajectory. (d) Heatmap of enriched TF motifs along cell clusters defining the JG trajectory. 100

Figure 5.10 **Enrichment of genomic functional classes in marker peaks along the JG differentiation trajectory**. Individual cell clusters along the differentiation trajectory leading to JG cells display differential enrichment of genomic classes including: 3' UTR, promoter proximal, promoter core, intron, intergenic, 5' UTR, exon, and enhancers. Bars for each class represent the observed proportion of regions defined as marker peaks for individual clusters relative to the expected proportion based on the number of bases defined as a particular genomic functional class. 101
Figure 5.11 **Positive transcription factor regulators of the JG differentiation trajectory** . TFs whose gene expression and motif enrichment are positively correlated indicate putative drivers of differentiation. 102
Figure 5.12 **Transcription factors enriched along the JG differentiation trajectory act as early to middle to late acting factors**. TFs that are enriched in clusters along the developmental trajectory of JG cells early (a), middle (b), or late (c). (b) **Bold** TFs are positive TF regulators *Italicized* TFs are known regulators of renin expression. JG: juxtaglomerular 103
Figure 5.13 **Transcription factors differentiating JG cells from mature SMCs are preferentially enriched in late differentiation trajectory clusters**. (a) Mef2c (and Mef2a (b), Mef2b (not shown), Mef2d (not shown)) are enriched in late time point clusters that contain and ultimately form mature JG cells. Zfp283 (c) is enriched middle to late along the differentiation trajectory. Tead3 (d) differentiates JGs from SMCs but is generally equally enriched across development. Stat5b (e) is positively enriched throughout the middle and late clusters along the JG differentiation trajectory. Bcl11a (f) and Bcl11b (g) differentiate JG cells from SMCs but are overall lowly enriched in all clusters. Nr2f6 (h) and Nr2c1 (i) are enriched in late time point clusters including JG cells. Rarg (j) is enriched in late developmental time points including JG cells. JG: juxtaglomerular 104
Figure 6.1 **UMAP visualization of scATAC-seq datasets using region-set2vec successfully separates single cells into biologically meaningful clusters.** (a) Simulated bone marrow dataset with a coverage of 2500 fragments per cell [455]. (b) Mouse Foxd1+ progenitors from four developmental time points: E12, E18, P5, and P30. 106

List of Tables

Table 1 **Step-by-step guides** 17
Table 2 **Raw ATAC-seq data processing pipelines** 19
Table 3 **Quality control tools** 20
Table 4 **Peak calling tools** 21
Table 5 **Tools to investigate differentially accessible regions** 21
Table 6 **Motif enrichment and transcription factor footprinting tools.** . 22
Table 7 **Tools to investigate nucleosome positioning.** 23
Table 8 **Tools to investigate region enrichments** 24
Table 9 **Tools for single cell ATAC-seq data processing** 26

Abstract

An ultimate goal of biology is the understanding, at a fundamental level, of the function of the cell. What factors shape cell identity and function and how does an organism or individual cell control the timing, development, and activity of its genome? A cell's identity can change over time, based on genetic and epigenetic signals, its spatial context in an organism, or due to internal or external stimuli. Genomics and epigenomics seek to uncover these signals by measuring molecular profiles of RNA [1–4], DNA [5–8], protein [9, 10], epigenetic modifications [11–13], or chromatin accessibility and conformation [14–18]. Open chromatin and gene expression assays with their corresponding computational tools enable the deconvolution of complex samples, the identification of rare or novel cell types and regulatory elements, and of the interactions between DNA and chromatin-interacting proteins [19–24]. We sought to evaluate the status of the computational infrastructure to enable these sorts of analyses, address unmet needs in the field, and apply this knowledge and expertise to investigate questions of development in a rare kidney cell (renin or juxtaglomerular cells) that is integral for maintaining homeostasis. To fulfill this goal, we evaluated current methods for open chromatin analysis, developed computational pipelines to analyze bulk ATAC-seq, nascent RNA-seq, and applied an integrated analysis of scATAC-seq and scRNA-seq to uncover the regions and factors driving the differentiation of renin cells in developing mouse kidneys. This work led to the novel finding of the importance of the MEF2 family of transcription factors being primary drivers of renin cell differentiation.

1 Introduction

1.1 Chromatin

1.1.1 Chromatin overview

Chromatin is dynamic and changes in chromatin accessibility to various transcription factors and remodeling complexes reflect changes in transcriptional activity of the cell. This accessibility is in essence a measure of the degree to which nuclear proteins and molecules are able to interact with the underlying DNA. Chromatin is itself tightly regulated to ensure proper function of DNA processes. This regulation can occur from the level of individual nucleosomes, to general DNA accessibility, up to and including higher-order structures of chromatin. Thus the organization of accessible chromatin across the genome represents the set of possibly interacting regulatory elements and chromatin binding factors that regulate gene expression [25, 26]. Factors that interact with chromatin are responsible for regulating this structure and can act both directly on chromatin conformation as well as indirectly dependent on that same conformational structure. Therefore, a complex system emerges whereby changes in chromatin structure affect which chromatin regulators are able to bind and chromatin regulators in turn affect the chromatin structure [27, 28]. While the genome is the same, cells execute unique functions based on specific gene expression patterns under regulation by the topological organization in the nucleus [29–32]

1.1.2 Chromatin structure

Chromatin structure enforces broad and significant effects on essentially all DNA based processes from kinetochore and centromere formation to DNA repair, replication, and transcription [[33]; [34]; Venkatesh2015]. Furthermore, beyond the underlying genetic sequences themselves, chromatin structure can also be inherited [34]. Chromatin is organized broadly into regions of highly condensed, transcriptionally silent areas mainly present at pericentric regions and telomeres known as heterochromatin versus euchromatin, which includes less condensed, gene-dense, actively transcribed regions [35–39].

Restructuring transcriptional networks is the method by which cells undergo development and respond to changes in their environment without changes to underlying DNA sequences. These primarily epigenetic changes occur during cell division and differentiation and may be partially controlled through changes in chromosomal conformation. Historically we viewed gene expression as occurring from a linear arrangement of DNA sequence encoding genes in a one-dimensional state. However, we now know that three-dimensional interactions are often

necessary and required to bring regulatory elements and chromatin-bound factors together with target genes to regulate expression. In this updated view, transcriptional activity and regulation involves the interplay between transcriptional complexes involving TFs [40, 41] and other regulators [42, 43] brought together through DNA looping [44, 45].

Chromatin's underlying structure is composed of approximately 147 bp of DNA wrapped around repeating units of eight histone proteins which together form a nucleosome [46–49]. The distribution of these nucleosomes is not uniform across the genome [50, 51]. Nucleosomes tend to be densely arranged in facultative and constitute heterochromatin, but are depleted at regulatory regions and actively transcribed gene bodies [50–53].

1.1.2.1 Nucleosomes The location of nucleosomes along chromatin serves as one of the determinant factors controlling accessibility of DNA to interacting proteins including transcription factors [54, 55]. The nucleosome comprises the major subunit of eukaryotic chromatin. Its core is comprised of approximately 147 bp of DNA wrapped 1.65 times around a histone octamer composed of two of each of the four core histones, H2A, H2B, H3, and H4 [46, 48, 49, 53, 56, 57]. The length of the DNA that wraps around a nucleosome core is tightly conserved with the length of DNA between nucleosomes variable and subject to different chromatin regions, different cell and tissue types, and different species [58]. These differences directly affect the activity of critical processes of cellular regulation by controlling which proteins have access to specific DNA regions. Assaying regions of open chromatin therefore provides a direct measure of which regions of a particular cell or groups of cells are accessible to cellular machinery.

1.1.2.2 Histones Histones are the molecular units that comprise nucleosomes and work to package and organize DNA in the nucleus [39]. The core histones that comprise a nucleosome include: H2A, H2B, H3 and H4 [57]. There is a wide range of post-translational modifications of histones that affect the function of each individual subunit and the accessibility of nearby DNA [59–61]. Furthermore, multiple variants exist in each of the core histone subunit families that play different roles in development, epigenetic regulation, and localization [39, 62]. These modifications may serve as epigenetic indicators of chromatin states and play important roles in determining or indicating the activity of regulatory elements [32].

For example, the histone variants H2A.Z and H3.3 can influence nucleosome turnover rates and are enriched at TSSs and enhancers [63, 64]. Furthermore, specific histone post-translational modifications are indicative of various regulatory elements. Nucleosomes downstream of active promoters are enriched for H3K4me3 and H3K27ac as are active enhancers (Fig. 1.1)[65–70].

Enhancers contain H3K4me1 whether they are inactive, poised, or active [67, 71]. Repressed or inactive enhancers include H3K27me3 and dense nucleosome assemblies (Fig. 1.1)[69]. To distinguish between active and poised enhancers, poised enhancers have H3K27me3 in place of H3K27Ac and reduced chromatin accessibility (Fig. 1.1) [69, 72–74]. Poised enhancers represent a possible epigenetic priming mechanism [75], are evolutionarily conserved during embryonic development, and appear to be a controlling mechanism of lineage specificity [74, 76].

H3K9ac	Active promoters and enhancers
H3K14ac	Active transcription
H3K4me3/me2	Active promoters and enhancers
H3K4me1	Enhancer-specific
H3K27ac	Enhancer-specific
H3K36me3	Active transcribed regions
H3K27me3/me2/me1	Silent promoters
H3K9me3/me2/me1	Silent promoters

Figure 1.1: **Histone modifications are indicative of promoter and enhancer activity**.

1.1.3 Regulatory regions

Regions of open chromatin represent areas of DNA accessible to transcriptional machinery. Thus, cellular identity and function is at least partially defined by what regions of chromatin are open and the corresponding regulatory elements present. Both genes and non-coding regions of the genome in these open areas participate in the expression and activity of proteins. Regulatory regions are found primarily in these accessible regions and the identification of cell-specific regulatory elements is critical to understand cell identify and function. While less than 2-3% of the genome encodes for amino acids in proteins, 90% of regions bound by a TF are found in the remaining regions of open chromatin [[51]; Cao2015]. Identifying open chromatin provides the opportunity to identify not only regulatory regions but binding sites of chromatin-interacting proteins that act to regulate gene expression. These regulatory regions are composed of several general classes including: promoters, enhancers, insulators and silencers. Both promoters and enhancers may initiate transcription but only at gene promoters are the resultant transcripts stable [70, 77–79]. Promoter activity is itself modulated by input from enhancers, with both mediated by transcription factors and transcriptional cofactors [70].

1.1.3.1 Promoters Promoters are genomic regions located near gene transcription start sites (TSSs) [70, 80, 81]. They may be further broken down into core and proximal promoters.

Core promoters are regions nearest TSSs (~50bp upstream and downstream of the TSS) where transcription machinery assembles, including the pre-initiation complex containing Pol II and general transcription factors (GTFs) with active cores devoid of nucleosomes [40, 70, 82–86]. The proximal promoter is located around 250 bp upstream of the TSS and contains transcription factor binding sites to regulate the corresponding gene's expression [70].

Within core promoters, there are several universally employed motifs with fixed positioning, including the well-known TATA-box motif [86, 87], the imitator motif [88, 89], TFIIB recognition elements [90, 91], or downstream core elements [92]. This enables the classification of core promoters into three general types. The first includes core promoters with imprecisely positioned nucleosomes and clear initiation patterns in addition to motifs for TATA-box and initiator [93, 94]. They often represent genes active in terminally differentiated cells [95]. The second type of core promoter is found at housekeeping genes which are broadly expressed. They tend to have less precise initiation patterns but with precisely positioned nucleosomes [93, 96, 97]. The third type resembles the promoters found at housekeeping genes but are in poised states, as indicated by H3K4me3 and H3K27me3 histone marks, suggesting they are important for directing cell lineage and terminal differentiation determination [98, 99].

1.1.3.2 Enhancers While core promoters are sufficient to initiate transcription, they generally have low basal activity which can be modulated by distal enhancers [86, 100, 101]. While the distinction between enhancers and proximal promoters is minimal, enhancers are distinguished by being distal to core promoters and can act independent of distance and orientation [70, 86, 101]. They are further distinguished from promoter elements by enrichment of H3K4 methylation, H3K27ac, and the presence of histone variant H2A.Z [66, 102–105]. They are present in regions of open chromatin and are activated by the binding of transcription factors and cofactors [70]. These recruited factors can interact with promoters to increase initialization or stabilization of transcriptional machinery through three-dimensional organization of chromatin [100, 106–110]. They can be located in *cis* or *trans* to genes they regulate [45, 81, 111, 112]. Enhancers are essential for controlling the specificity of gene expression with different enhancers active in different cell types and tissues. The combination of active enhancers in a cell controls the expression of cell identity genes and represent the primary indicator of cell specificity [81, 113–115]. Furthermore, clusters of enhancers in close proximity bound by TFs that are associated with cell specificity have been identified that are termed super-enhancers, although their function as distinct functional units is unclear [116, 117].

1.1.3.3 Insulators For enhancers to interact with target gene promoters, they must be within regions that are topologically accessible to each other. Insulators are elements that can set the boundaries of these topologically accessible regions and are defined by the binding of CTCF [118–120]. They function by ensuring the correct distal elements have access to the right promoter elements [121–123]. The region bound by CTCF-bound insulators is highly accessible but surrounded by dense arrays of stable nucleosomes [32].

1.1.3.4 Silencers Silencers are regulatory regions which repress gene activity by modulating chromatin looping at the target promoter [124, 125], recruiting repressive transcription factors [126, 127], or through recruiting histone writers to apply repressive chromatin marks [128]. While their genomic localization is similar to enhancers and they can act independent of orientation and distance to promoters, they result in decreases in target gene transcription in contrast to enhancers [129].

1.1.4 Transcription factors

Transcription factors (TFs) are proteins that bind specific DNA sequences via a DNA-binding domain and can regulate transcription [130–132]. TFs do so via the recruitment of RNA polymerase II or transcriptional cofactors through a transactivation domain [133, 134]. The binding and activity of TFs is a dynamic process with competition between histones and chromatin-interacting proteins to affect nucleosome occupancy and control accessibility of DNA [[135] ;[136]; Felsenfeld1996; [137]], Furthermore, the chromatin accessibility landscape is itself dynamic, with active chromatin remodeling modulating nucleosome turnover and affecting the ability of TFs to bind and recruit cofactors and distal regulatory elements [138–141]. Although the majority of TFs require accessible DNA to bind, pioneer TFs are thought to bind directly to nucleosomal DNA or are the first factor that binds following the establishment of open chromatin [136, 142–148].

1.2 Assays for measuring open chromatin

The ability to determine a cell's chromatin landscape is essential for understanding the regulatory processes driving cellular function and identity. Open chromatin assays identify DNA regions that are accessible to external factors (e.g. TFs, chromatin remodellers, RNA pol II), and these regions have been shown to correspond to regulatory elements, including promoters, enhancers, and others [26, 51, 149–151]. Regulatory element activity varies spatially, temporally, and between cell-types to influence the binding of transcription factors and the expression of target genes [149, 150]. Studying activity of regulatory elements promises

to increase our understanding of the fundamental biology of gene regulation, and its influence on human health and disease [152–160]. Measurements of open chromatin in both bulk and single cell populations act as proxies for TF-binding signals providing further insight into epigenetic regulation [32, 51].

1.2.1 MNase-seq

With the central role of histone proteins in regulating chromatin accessibility, MNase-seq is a technique developed for assaying nucleosome occupancy [58, 161, 162]. MNase-seq derives its name from the use of a non-specific micrococcal nuclease (MNase) derived from the bacteria, *Staphylococcus aureus*. This nuclease has a strong preference for cleavage in non-nucleosomal regions, thus providing a means of enrichment of accessible chromatin [58]. If a region of DNA is bound to histones or a transcription factor, MNase is unable to bind and cleave that region. A limitation of MNase-seq for assaying chromatin broadly is that it is primarily focused on identifying fragments of DNA bound directly by the histones or other chromatin binding proteins. Therefore, MNase-seq identifies regions of DNA which are transcriptionally inaccessible, and is not directly applicable for investigating epigenetic regulation of genes [163].

1.2.2 DNase-seq

DNase-seq uses the DNaseI enzyme to digest regions of chromatin unprotected by bound proteins, leaving behind accessible regions that are known as DNase I hypersensitive sites (DHSs) [164–166]. DNA regions tightly wrapped around nucleosomes or bound in higher-order structures are effectively protected from digestion [166]. By isolating DHSs followed by high-throughput sequencing, DNase-seq enables the whole-genome interrogation of open chromatin and the identification of active gene regulatory elements. While all chromatin assays have some level of bias dependent on the cleavage process employed, DNase-seq shows an increased preference for signal at promoters, but maintains a comparatively low signal-to-noise ratio enabling identification of all classes of regulatory elements [167, 168].

1.2.3 FAIRE-seq

FAIRE-seq was designed to provide a straight-forward method for isolating nucleosome-deprived DNA from human chromatin [168, 169]. It relies on the use of formaldehyde to crosslink target proteins with DNA followed by sonication or pulverization and phenol-chloroform extraction to separate DNA that is crosslinked or not. The resulting non-crosslinked DNA is sequenced and provides a direct measure of open chromatin. However, relative to

competing open chromatin assays, FAIRE-seq requires higher input of cells [170]. Its reliance on crosslinking also means that even transiently bound proteins may reduce the amount of non-crosslinked DNA obtained [171]. FAIRE-seq also suffers from a lower signal-to-noise ratio compared to competing approaches meaning identifying signal compared to background can be more challenging [168]. It also tends to be biased to non-promoter regions of the genome which can be leveraged as an advantage if distal regulatory elements are of greater interest [167, 168, 172].

1.2.4 ATAC-seq

The most recently developed approach for assaying bulk open chromatin is ATAC-seq [170, 173, 174]. ATAC-seq dramatically improved the efficiency in cost, time (hours versus days), and required amount of sample over previous chromatin assays [170]. ATAC-seq relies on the activity of a hyperactive Tn5 transposase [170, 175]. This transposase is leveraged, through a process known as tagmentation [176], to fragment the genome while simultaneously inserting sequencing adapters [170]. This enables the determination of open chromatin sequences without affecting the underlying chromatin structure prior to sequencing by avoiding high-salt conditions, sonication steps, or crosslinking. An early limitation to ATAC-seq was the presence of high amounts of mtDNA contamination, but both library preparation [174] and computational improvements [177] have addressed this limitation.

1.2.5 scATAC-seq

Further refinements in ATAC-seq protocols have led to the ability to assay open chromatin in single cells (scATAC-seq) [17, 178, 179]. These tools have uncovered the variability of accessibility among cells to establish distinct molecular states and the identification of rare cell populations [17, 18, 160, 180, 181]. Each of these methods utilizes cellular barcodes or indexes to link the source of sequenced reads back to individual cells following the previously established ATAC-seq assay [17, 178, 179]. The primary benefit of single cell methods compared to the previously described bulk assays is the identification of open chromatin in heterogeneous cell populations. This is particularly important when studying dynamic processes such as development or responses to stimuli where multiple cell subpopulations modulate their accessible chromatin differentially. The primary caveats to scATAC-seq over bulk methods includes increased time, starting material, cost, and the complexity of bioinformatic analysis.

1.3 Assays for measuring gene expression

Measurements of open chromatin provide direct evidence for epigenetic regulation of gene transcription. However, they are independent of direct measures of gene activity itself. Assays for measuring RNA abundance do and by integrating the two approaches we improve our understanding of the shared regulatory processes controlling cell identity and function.

1.3.1 RNA-seq

RNA-seq includes any method for assaying RNA abundance in large quantity using high-throughput sequencing [182–184]. It leverages the reverse transcription of cellular RNA, ligation of sequencing adapters, and deep sequencing of the resultant libraries. RNA-seq can therefore measure the steady-state abundance of total RNA, mRNA, or small RNAs including rRNA [185], miRNA [186], or tRNAs [187]. Sequenced RNA can be then be used to describe changes in gene expression over time [188, 189], between experimental conditions [190], or the identification of gene fusions [191] and alternative splicing events [192]. As one of the earliest applications of next-generation high-throughput sequencing, RNA-seq represents well-established and cost-effective protocols for assaying gene expression in cells.

1.3.2 scRNA-seq

Similar to differences between bulk and single-cell chromatin assays, bulk RNA-seq masks individual contributions of cellular subpopulations as aggregate signals of expression [22, 193, 194]. To address limitations of bulk RNA-seq, scRNA-seq methods have undergone rapid development and adoption [3, 4, 194–197]. With these technologies, deconvolution of heterogeneous cell populations can be performed [198–200]. Cells' expression profiles can be categorized by type, spatial organization, and through stochastic differences in gene regulatory networks across time [194, 200, 201].

1.3.3 Nascent RNA-seq

Both bulk and scRNA-seq represent steady-state levels of gene expression, even where differences between individual cells are present. As cells are dynamic entities, response to both internal and external stimuli is rapid and on much shorter timescales than can be captured with non-nascent approaches. Nascent RNA sequencing methods address this limitation in several ways. Instead of measuring a snapshot of stable mRNA accumulation and turnover, nascent RNA sequencing directly measures transcription, the orientation of transcripts, and can capture unstable, short-lived transcripts too [202–204]. These latter features are significant as bidirectional transcription is captured at promoters and enhancers. Therefore, nascent

RNA sequencing provides an independent method to directly identify active enhancers in cells [[205]; [206]; [207]; [208]; [79]; [209]; [210]; [211]; Wang2019].

1.4 Computational challenges of genomic and epigenomic analysis

With the rapid development of both bulk and single-cell assays for gene expression and chromatin, corresponding computational methods have lagged the proliferation of available data. While the more established bulk RNA-seq assays have enjoyed a robust development of analysis techniques [212], bulk ATAC-seq has only recently experienced settled field-standard analyses [177, 213, 214]. Conversely, with the more recent advent of single-cell RNA-seq, nascent RNA-seq, and single-cell ATAC-seq, computational methods for standard and efficient analysis is again an active and needed area of development [194, 215, 216].

1.4.1 Bioinformatic analysis of single-cell genomics

Single-cell genomic assays resolve limitations of bulk approaches, but suffer from increased complexity in computational analyses. Due to the sparse and high-dimensional data from single-cells, new techniques became necessary to properly normalize and evaluate single-cell reads. In scRNA-seq, dropouts, or transcripts with zero reads either from biological or technological reasons become a concern as are batch effects, increased heterogeneity, and more complex distributions of expression [217, 218]. In scATAC-seq, data is sparse since diploid organisms only carry two copies of DNA, and thus any measure of chromatin accessibility from a single cell is binary [214, 216]. This sparsity is exacerbated in scATAC-seq due to there being hundreds of thousands of possible regulatory regions compared to the relatively compact twenty thousand genes in a scRNA-seq experiment. Therefore, scalable computational approaches that address the size of data, data sparsity, batch effects, normalization, visualization, dimensionality-reduction, and clustering approaches are necessary [181, 216, 218–224]. Finally, combining both single-cell chromatin accessibility with single-cell transcriptomics can uncover the interplay between regulatory elements, transcription factor activity, and their impact on expression [32, 216, 224, 225].

1.5 Renin cell development

With the challenges posed by renin cell rarity and intractability to *in vitro* methods, single-cell omics provide the necessary tools to begin uncovering how these cells are formed and regulated. We sought to utilize these approaches to answer the following questions. Can we identify this rare cell population through enrichment of lineage-tracing models of renin cell progenitors? Once identified, can we define a trajectory of early progenitor subpopulations before arriving

at the mature renin-producing cell? What genetic and epigenetic changes occur during this development, and can we identify TFs potentially involved in regulating any observed epigenetic differences? Are TFs expressed at different times or in different subpopulations? Are there accessible binding sites for putative effector TFs present in our cell population of interest? By combining scRNA-seq and scATAC-seq, we finally have the tools to begin answering these questions.

1.5.1 Renin cell overview

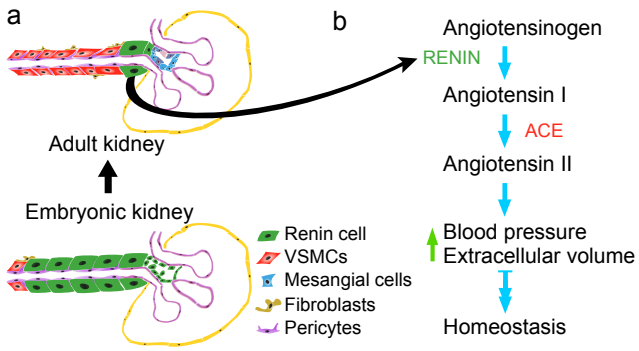

Figure 1.2: **Distribution of renin cells in development and the role of renin in homeostatic control**. (a) Renin cells line afferent arterioles during embryogenesis before differentiating into vascular smooth muscle cells (VSMC) and mesangial cells with mature renin-expressing cells restricted to the juxtaglomerular region in adult mammals. (b) Renin converts angiotensinogen into angiotensin I to initiate homeostatic balance.

Renin cells are critical for survival by maintaining homeostasis through the release of the hormone-enzyme renin in response to minute changes in blood pressure [226–228]. These renin-producing cells are restricted in adult mammals along the walls of renal arterioles near the entrance to the glomeruli, and are therefore known as juxtaglomerular (JG) cells [228–230] (Fig. 1.2a). Renin release initiates a cascade that produces angiotensin II, leading to vasoconstriction and blood pressure increase (Fig. 1.2b). Not only do renin cells play this vital role, they are also progenitors for multiple additional cell types that retain the memory of the renin phenotype and are able to restore this phenotype to produce renin under stress [228, 230]. Despite the clear importance of these cells for organism health, we still do not fully understand their development. Building off past knowledge, our goal is to define the mechanisms that govern the identity and plasticity of renin-expressing JG cells.

Renin cell research is made difficult through a number of challenges. First, renin cells are incredibly rare, accounting for 0.01% of the kidney cell mass [231]. Second, they are incredibly challenging to isolate and stop producing renin after only 48 hours in culture [231]. Despite these issues, a preponderance of evidence has grown to describe renin-cell physiology. They are known to originate from Foxd1 positive (Foxd1+) mesenchyme cells in the kidney and are

themselves progenitors to smooth muscle cells, pericytes, mesangial and tubular cells [230–235]. Studies in lineage-tracing mouse models, transcriptomic and epigenomic analyses have also begun to expand our understanding of the development and control of renin-expressing cells.

1.5.2 Current knowledge of renin cell regulation

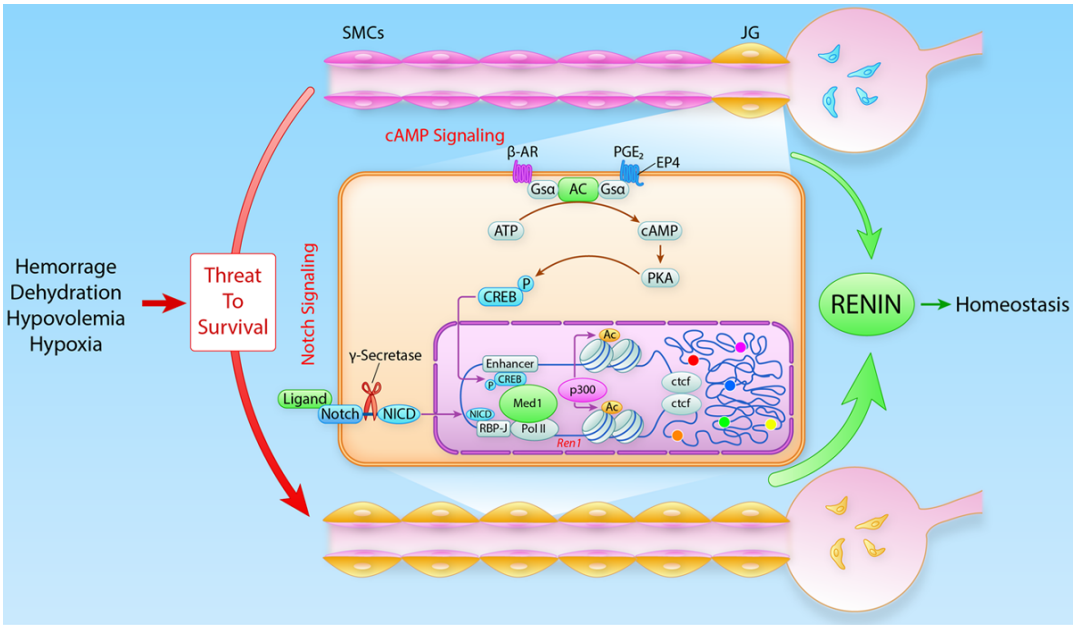

Figure 1.3: **Super-enhancers (SEs) act as chromatin sensors that control the identity and memory of renin cells to maintain homeostasis.** Schematic summarizing the main signaling pathways, and chromatin changes involved in the maintenance of juxtaglomerular (JG) cell identity and reacquisition of the renin phenotype by smooth muscle cells (SMCs) in response to physiological demands. Activation of the cAMP or the Notch pathways leads to profound epigenetic changes at the renin locus regulatory region characterized by deposition of acetylation of lysine 27 of H3 by p300, sliding of nucleosomes, and opening of chromatin, which facilitate the access of numerous transcription factors including but not limited to Med1 (Mediator complex 1), Creb (cAMP-responsible element binding protein) 1, and RBP-J (recombination signal binding protein for immunoglobulin kappa J region). Loop formation is maintained by Ctcf (CCCTC-binding factor). The colored dots indicate the presence of additional SEs throughout the genome that also regulate renin cell identity. β-AR indicates beta adrenergic receptor; AC, adenylate cyclase; CREBP, phosphorylated cAMP-responsive element binding protein; EP4, Prostaglandin E2 receptor 4; Gsα, activating G-protein–coupled subunit; NICD, Notch intracellular domain; PGE2, prostaglandin E2; PKA, protein kinase A; Pol II, RNA polymerase II; and RBP-J, recombination signal binding protein for immunoglobulin kappa J region. Illustration credit: Ben Smith. Data derived from Martinez et al. [228]. Included by permission from Maria Luisa S. Sequeira-Lopez and R. Ariel Gomez. Renin Cells, the Kidney, and Hypertension. Circulation Research. Volume: 128, Issue: 7, Pages: 887-907, DOI:10.1161/CIRCRESAHA.121.318064. and the publisher, Wolters Kluwer Health, Inc. Please contact permissions@lww.com for further information.

Major efforts to elucidate determinants of renin cell identify have uncovered a number of important pathways and genomic regions integral to renin-expressing cells. The cAMP pathway has been shown to stimulate renin gene transcription and subsequent release (Fig. 1.3) [236–238]. The renin gene contains a cAMP responsive element where the histone acetyl

transferases CBP/p300 can bind to regulate renin expression [239–241]. Additionally, the final common effector of the Notch signaling pathway, RBP-J, is necessary to maintain renin expression and modulates the plasticity of SMCs and mesangial cells to restore renin expression [230, 242–244]. RBP-J also regulates AKR1b7 which is co-expressed with renin and serves as an additional marker of mature renin cells [230, 245]. Understanding the epigenetic changes that occur to regulate the renin phenotype is ongoing. More center efforts to uncover these changes identified a set of super-enhancers unique to renin cells [228]. The primary super-enhancer was found just upstream of the renin gene (Ren1) and is thought to be responsible for the restoration of renin phenotype in renin cell descendants [228]. Despite this knowledge of renin control, we are only beginning to undercover the epigenetic changes that occur along the differentiation trajectory of renin cells. An improved understanding of the dynamic genetic and epigenetic changes that occur in renin differentiation is necessary to better understand kidney pathologies and the effects of therapeutic targeting in cardiovascular disease.

2 Analytical approaches for ATAC-seq data analysis (modified from [213])

To identify gaps in the field of ATAC-seq data analysis, we performed a comprehensive survey of the most accepted approaches to analysis and the tools and infrastructure available. Based upon these findings, we realized there existed a substantial gap in easy to adopt, well-documented, robust and reproducible pipelines for ATAC-seq analysis that we later sought to address.

2.1 Introduction

As our understanding of gene regulation has improved, so has our awareness of the increasingly complex chromatin landscape that governs that regulation. Assays to better evaluate this landscape have been rapidly developed and improved, and the Assay for Transpose Accessible Chromatin using sequencing (ATAC-seq) has become a common first step for studying gene regulation. ATAC-seq interrogates *chromatin openness*, or *chromatin accessibility*, similar to earlier assays such as DNase-seq, MNase-seq, or FAIRE-seq [246, 247]. These assays identify DNA regions that are accessible to external factors, which have been shown to correspond to regulatory elements, including promoters, enhancers, and other types of elements [26, 51, 149–151]. Activity of regulatory elements varies spatially, temporally, and among cell-types to influence the binding of transcription factors and the expression of target genes [149, 150].

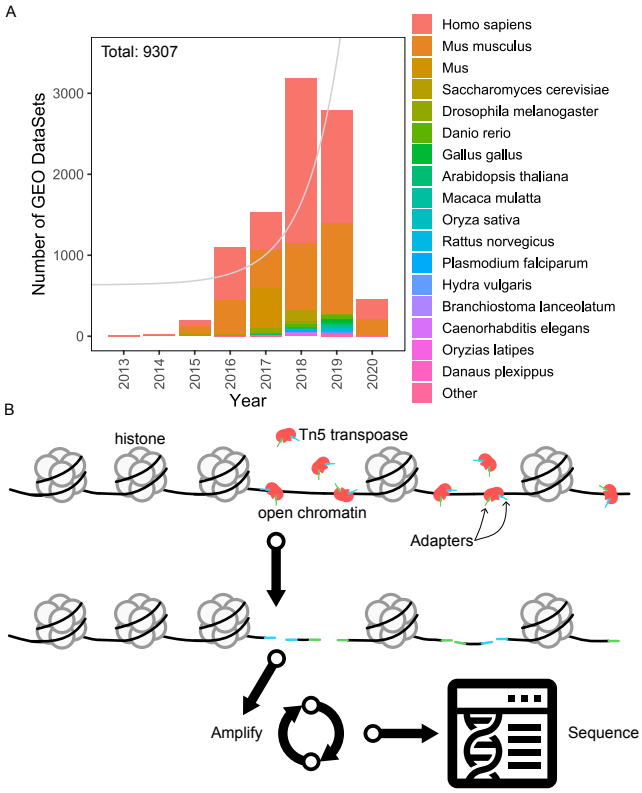

Figure 2.1: **ATAC-seq is a rapidly growing method for open chromatin analysis.** (a) Increasing prevalence of 'ATAC-seq' DataSets in the Gene Expression Omnibus (GEO) (Color = Species, Gray line = fitted exponential growth model) (b) Generalized ATAC-seq library prep protocol

Studying activity of regulatory elements promises to increase our not only understanding the fundamental biology of gene regulation, but also its influence on human health and disease [152–160].

ATAC-seq has been adopted rapidly in the scientific community, with the number of studies using ATAC-seq approaching 10,000 in just a few years (Fig. 2.1A). The primary factor driving this adoption is efficiency, as ATAC-seq dramatically improved the efficiency in cost, time, and required amount of sample over previous similar assays [170]. ATAC-seq relies on the activity of a hyperactive Tn5 transposase [170, 175]. This transposase is leveraged, through a process known as tagmentation [176], to simultaneously fragment the genome while inserting sequencing adapters [170]. These sequences can be PCR amplified and then sequenced using 2-4 orders of magnitude fewer cells, fewer protocol steps, and less time than analogous assays (Fig. 2.1B) [170, 248]. Protocols for ATAC-seq have improved since it was first introduced in 2013 [170, 173], for example, with improved removal of contaminating mitochondria DNA [174, 249] and extension to single cells [17, 178, 250]. As the protocol has developed and increased in popularity, analytical approaches have also been multiplying rapidly. Here, we provide guidance for both novice and experienced analysts on the advantages and limitations of ATAC-seq analysis pipelines, methods, and tools.

2.2 Fundamentals of ATAC-seq Data Analysis

A typical ATAC-seq analysis can be divided into two major components: 1) general processing of raw sequencing reads, which produces intermediate outputs like annotated peak calls; and 2) detailed downstream analysis, which is more specific to a particular biological question (Fig. 2.2). In general, the first step is universal to all downstream analysis types, whereas the second step then requires more specialized software.

2.3 Alignment, adapters, and mitochondrial reads

Analysis of ATAC data typically starts by processing raw sequences through a series of pipeline steps into outputs amenable for detailed biological questions (Fig. 2.2). A generalized workflow includes the following: First, reads are screened for quality, then adapter sequences are removed, and finally the reads are aligned to a reference assembly. After alignment, many pipelines are equipped to handle high mitochondrial DNA content, because ATAC-seq libraries are prone to high levels of mitochondrial DNA, which is typically considered undesirable. While recent protocol adaptations have succeeded in reducing mitochondrial DNA using optimized reagents [174, 251] or molecular biology techniques [249], many pipelines

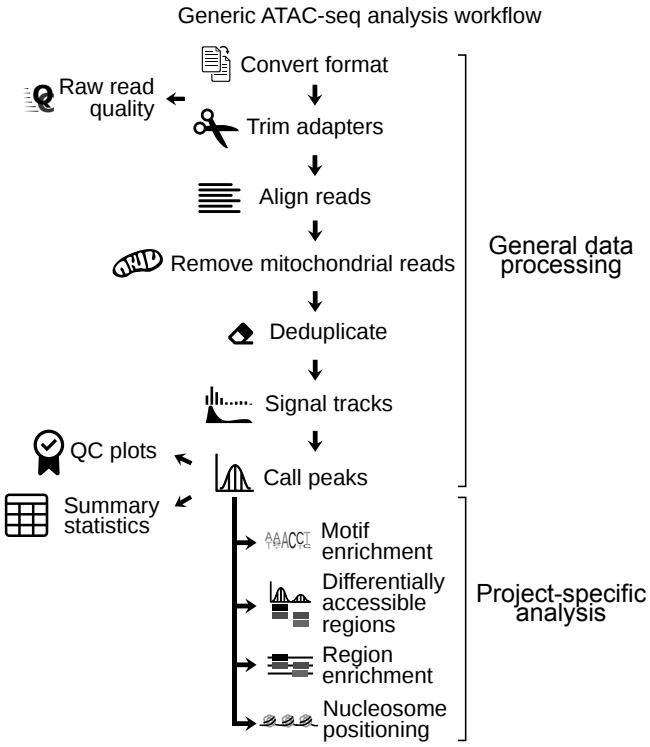

Figure 2.2: **ATAC-seq general workflow.** Raw reads are processed through a series of steps to produce uniform intermediate results, which can then be further analyzed with more analysis specific to a biological research question.

address this computationally by filtering out mitochondrial sequences. These sequences are removed through either sequential alignments to mitochondrial DNA before genomic, through removal of mitochondrial DNA from genome-wide genomic indices, or through blacklists of mitochondrial DNA after alignment. In our work, sequential alignment is the most accurate and computationally efficient way to eliminate mitochondrial contaminants – and it also allows for later analysis of mitochondrial reads [213].

2.4 Removing duplicates

Following adapter removal and alignment, pipelines remove read duplicates, although typical computational strategies may be overzealous in this approach if using only single-end sequencing data since there is only a single end to compare. Single-end sequencing also provides less information as it reduces the ability to identify PCR duplicates, which are typically removed, and it lacks the ability to identify both ends preventing the identification of where the transposase inserted. For these reasons, it is recommended to use paired-end ATAC-seq data when possible. After alignment and duplicate removal, low-quality, multi-mapping, or unmapped paired reads also typically get removed from downstream analyses.

2.5 Generating signal tracks

Once reads are aligned and filtered, they are shifted to accommodate the mechanics of transposase Tn5 activity [170, 175, 176]. When the Tn5 tranposase interacts with DNA, it effectively occupies about 9bp of DNA and introduces the sequencing adapter at the 5' end of the interaction site. The Tn5 adapters are inserted in a staggered manner to the 5' ends of target sequence strands with a 9 bp gap between them [170, 175, 176].This means the center of the Tn5 binding is actually 4 bp to the right of the edge of positive strand reads, or 5 bp to the left on negative strand reads. This shifting is intended to identify the center of the locus where Tn5 interaction occurred. An alternative approach is to account for the 9 bp size of the transposase binding event by mapping the reads as 9 bp insertion events instead of at nucleotide resolution. In either case, mapped reads are then transformed into signal tracks for visualization and further data analysis.

2.6 Peak calling

As the goal of ATAC-seq is the identification of regions of accessible chromatin, and by proxy, regulatory elements and sites of transcription factor binding, we must next identify those regions of interest. To do this, we identify areas of the genome that are enriched for aligned reads. These regions are identified and visualized as peaks. Calling peaks therefore represents the identification of regions of concentrated ATAC-seq signal which indicate regions of open chromatin. Peak calling necessitates choosing an appropriate peak-calling algorithm or tool that balances sensitivity and specificity of called peaks. User-defined settings can widely influence the number, width, and confidence of identified peaks [252]. Following the identification of peaks, they are typically broadly annotated into genomic partitions including known features such as promoters, exons, introns, or 3' and 5'UTR among others.

Peak calling is typically the end of the general data processing pipeline that considers each sample independently. With signal tracks and called peaks for each sample, analysts are prepared for downstream analyses using more specialized analysis approaches that depend on specific user-defined biological questions.

2.7 Downstream analysis

For detailed downstream analysis, the data is generally integrated across samples. These analyses include differential accessibilty analysis, motif analysis, footprinting, and peak enrichment analysis. Because these analyses are more specific to particular biological questions, they are not typically performed by general-purpose ATAC-seq pipelines and must be manually

set up for each study. Therefore, only a subset of these analyses will be relevant for a particular analysis, which should be determined before investing significant effort in a particular tool. We describe these analysis types in more detail in the next section.

2.8 Survey of Tools for ATAC-seq Analysis

Here, we present a survey of tools divided into classes based on their primary goal. This includes three classes geared toward the general ATAC-seq data processing: *Step-by-step analysis guides*, *Raw sequence pipelines and workflows*, and *Quality control* tools. The remaining tools are for more detailed downstream analyses, which we sub-divided into five categories: *Peak calling*, *Motif enrichment and footprinting*, *Nucleosome positioning*, *Differential accessibility*, *Region enrichment*, and *Single-cell analysis*. The advantages and disadvantages of the tools vary widely, and some are targeted for novices while others require an experienced analyst. Our survey provides an overview of each analysis type, along with a table of some characteristics of relevant tools, such as mode of operation, language, update frequency, and a link to more information.

Author	**Title**	**Notes**	**Update**
Yiwei Niu	ATAC-seq data analysis: from FASTQ to peaks	Blog style walk-through of generalized ATAC-seq data analysis.	2019
Steve Parker	BIOINF525 Lab 3.2	Minimal standard ATAC-seq analysis walk-through.	2016
Rockefeller University Bioinformatics Resource	Analysis of ATAC-seq data in R and Bioconductor	Bioconductor ATAC-seq analysis course.	2018
John M. Gaspar	ATAC-seq	Generalized ATAC-seq analysis walkthrough.	2019
Delisle L; Doyle M; Heyl F	ATAC-seq data analysis	Galaxy training walkthrough of generalized ATAC-seq analysis.	2020

Table 1: **Step-by-step guides**

2.8.1 Step-by-step Analysis Guides

For users who would prefer following a manual, stepwise procedure, several tutorials are available to walk a user through ATAC-seq data analysis (Table 1). These guides are a great starting point for an inexperienced user as they explain how each step is manipulating raw data towards the goal of called peaks and further analyses. Users are required only to be able to work at the command line and have experience installing prerequisites. Examples

include either formal classes available publicly (Steve Parker, Rockefeller University), training guides from public platforms [253], or guides from individual researchers sharing their own experiences (*e.g.* Yiwei Niu and John M Gaspar). These step-by-step guides are primarily educational tools and are not intended to be automatic, re-usable pipelines that can be easily deployed on many samples across multiple projects; for this application, users will be more interested in the reusable pipelines described next.

2.8.2 Raw Sequence Pipelines and Workflows

A more common need is a standardized pipeline to process raw data through fastq processing, alignment, peak calling, and signal track generation (Fig. 2.2). A number of raw data processing pipelines are available (Table 2). Many comprehensive pipelines now exist with different target audiences. Some pipelines are geared toward the bench biologist with graphical user interfaces, including both open-source (I-ATAC, GUAVA) and commercial options (Basepair). While the GUI may simplify things for some users, these tools tend have less documentation and also give less power to the user. The majority of raw data processing pipelines are executable at a command line interface (CLI). Among these pipelines, there is a wide range of possible pipeline end-points. Some pipelines are geared toward doing only universal analysis, ending at annotated peaks to provide a starting point for more detailed downstream analysis. Other pipelines include substantial cross-sample analysis after peak calling. To delineate this distinction, we have categorize pipelines into two groups: *entry-point* pipelines provide a series of outputs intended as the beginning of a user-controlled downstream analysis, while *end-point* pipelines are intended as a complete analysis, running integrated analysis internally.

Entry-point pipelines (AIAP, ENCODE, `PEPATAC`) are generally robust and reproducible to yield consistent processing of few to many samples. This goal necessarily excludes some downstream steps to improve efficiency, and for the fact that not all researchers may wish to do all analyses all the time. This is particular important if those additional procedures are not specific to the biological question being investigated. In that case, those additional procedures come at the increased cost of time and computational resources. All three of the entry-point pipelines include some level of shared and novel quality-control metrics to identify quality libraries with minimal project-specific analyses included.

The majority of the pipelines are end-point oriented, with substantial downstream processing following peak calling and signal track generation. The advantage of end-point pipelines is that they require the least additional effort for a complete analysis. These pipelines typically

Name	Language	Notes	Docs	Last Update	Citation
AIAP	Bash, R, Python	Optimized analysis with novel QC metrics	++	2019	Liu et al. (2019)
ATAC2GRN	Bash, Python	Parameter optimized ATAC-seq pipeline	+	2018	Pranzatelli et al. (2018)
ATAC-pipe	Python, R	"Analysis pipeline for ATAC-seq data including TF footprinting, cell-type classification, and regulatory network creation"	+++	2019	Zuo et al. (2019)
ATACProc	Bash, Python, R	Complete pipeline with additional downstream analyses included	++	2019	unpublished
Basepair	NA	Commercial. Web-based GUI for complete analysis	Unknown*	Unknown*	unpublished
CIPHER	R, Perl, Python	A data processing platform for ChIP-seq, RNA-seq, MNase-seq, DNase-seq, ATAC-seq and GRO-seq datasets	+	2017	Guzman and D'Orso (2017)
ENCODE	Python, Bash	Complete pipeline following ENCODE standards for ATAC/DNase-seq analysis	++	2020	unpublished
esATAC	R	Complete pipeline including downstream analyses	+++	2019	Wei et al. (2018)
GUAVA	Java, Python, R	GUI based complete ATAC-seq pipeline	+	2019	Divate and Cheung (2018)
I-ATAC	Java	GUI based interactive ATAC-seq pipeline	+	2017	Ahmed and Ucar (2017)
Nfcore-atacseq	Python, R	Complete pipeline build using Nextflow	+++	2019	Ewels et al. (2019)
PEPATAC	Python, R, Perl	Complete pipeline with unique analytical approaches and QC metrics	+++	2019	unpublished
pyflow-ATAC-seq	Bash, Python	ATAC-seq snakemake pipeline with included nucleosome positioning and TF footprinting	++	2020	unpublished
snakePipes ATAC-seq	Python	Workflow system including, but not limited to, ATAC-seq analysis	+++	2019	Bhardwajet al. (2019)
Tobias Rausch	Bash, R, Python	Complete pipeline with emphasis on downstream analyses	++	2020	Rausch et al. (2019)

Table 2: **Raw ATAC-seq data processing pipelines**

include the ability to incorporate sample structure (case vs. control) for differential analysis of accessible regions, transcription factor binding sites, or motifs. However, the cost of this convenience is a lack of customizability, as the exact downstream analysis may or may not match the requirements of a particular study, and the exact settings and assumptions must be taken into account. Furthermore, the increased complexity of pipelines that include numerous downstream analyses may waste analysis time and computational resources if that analysis is irrelevant for the question under investigation.

2.8.3 Quality control

Name	Languages	Notes	Docs	Reference
ATAqC	Bash; Python	Generate ATAC-seq specific quality control metrics.	+	unpublished
ATACseqQC	R	Provides ATAC-seq specific quality control metrics and transcription factor footprinting.	+++	Ou et al. (2018)
ataqv	C++; Bash	ATAC-seq QC and visualization.	+++	unpublished

Table 3: **Quality control tools**

Raw data processing pipelines have nearly universally adopted several standard quality control (QC) metrics. Briefly, these include QC of the raw and aligned sequence data, the distribution of aligned sequence fragments to confirm the presence of nucleosomes, measures of library complexity, the fraction of reads in peaks (FRiP), and the enrichment of reads at transcription start sites (TSS). Quality control tools are dedicated tools that provide these and more advanced QC metrics (Table 3). Advanced metrics include the enrichment of promoter signal relative to gene body, measures of the proportion of nucleosome free reads, and measures of signal to noise.

2.8.4 Peak calling

Comprehensive ATAC-seq pipelines typically employ one of just a few widely adopted peak callers, which include tools originally developed for ChIP-seq or DNase-seq experiments, such as F-Seq [254], MACS [255], or PeaKDEck [256]. There are also other options built specifically for ATAC-seq data, including Genrich [257] and HMMRATAC [258] (Table 4). The widely employed peak callers developed for ChIP-seq and DNase-seq experiments offer the advantage of years of demonstrated utility, support, and understanding of their strengths and weaknesses, but may neglect features of ATAC-seq data such as nucleosome positioning and transposase biases. Because ATAC-seq seeks to identify regions of open chromatin, the peak calling step is critical, so there will likely continue to be effort dedicated to improving peak calling tools and leveraging ATAC-specific data features to improve accuracy.

Name	Languages	Notes	Docs	Reference
F-Seq	Java	Can be used as general peak caller to identify regions of open chromatin.	++	Boyle et al. (2008)
Genrich	C	Peak caller for genomic enrichment assays with specific ATAC-seq mode.	+++	unpublished
HMMRATAC	Java	Identify nucleosome positioning and leverage ATAC-seq specific read outs to call peaks.	+++	Tarbell and Liu (2019)
Hotspot2	C++	Identify significantly enriched genomic regions.	++	unpublished
HOMER	Perl; C++	Suite of tools that include the ability to call peaks from DNA enrichment assays.	+++	Heinz et al. (2010)
MACS2	Python	Specifically designed for ChIP-seq but broadly applicable to any DNA enrichment assay to call peaks.	+++	Zhang et al. (2020)
PeaKDEck	Perl	Peak calling program for DNase-seq data.	+++	McCarthy and O'Callaghan (2014)

Table 4: **Peak calling tools**

2.8.5 Differential accessibility

Name	Languages	Notes	Docs	Reference
DAStk	Python	Identify changes in transcription factor activity by looking at changes in chromatin accessibility	+++	Tripodi et al. (2018)
diffTF	Python; R	Identifies differential transcription factors. Can operate in basic mode with just chromatin accessibility or in classification mode where it integrates RNA-seq.	+++	Berest et al. (2019)

Table 5: **Tools to investigate differentially accessible regions**

ATAC-seq peaks correspond to regions of open chromatin, which have been shown to identify regulatory regions. One of the most common analysis is to identify differentially accessible regions. Analagous to identifying differential expression between two sample types, differential accessibility can demonstrate how gene regulation is goverend in different biological settings. Typically, differential regions are identified by counting sequencing reads in individual peaks, and then using mainstream count-based statistical tests to assess for statistical differences. Most analysis uses popular R packages for count-based data, such as edgeR [259, 260], DESeq2 [261], or DiffBind [262]. While designed for other data types, like RNA-seq, because ATAC-seq data is count-based, the statistical assumptions are often transferable.

After identifying differentially accessible regions, we typically want to better understand what factors are acting at these regions. A common follow-up is to identify which transcription factors are also differentially active between scenarios (Table 5). To accomplish this, there are at least two tools optimized to work with ATAC-seq data to identify differential transcription factor activity. By incorporating chromatin accessibility information and reported transcription factor binding sites it becomes possible to identify differential TF activity [DAStk 263, diffTF, 264]. Should an experiment also include corresponding gene expression information, its possible to then classify differential transcription factors as activators or repressors [264].

2.8.6 Motif enrichment and TF footprinting

Name	Languages	Notes	Docs	Reference
BiFET	R	Identify overrepresented transcription factor footprints.	++	Youn et al. (2019)
BinDNase	R	Transcription factor binding prediction using DNase-seq.	+	Kahara and Lahdesmaki (2015)
CENTIPEDE	R	Transcription factor footprinting and binding site prediction.	++	Pique-Regi et al. (2011)
DeFCoM	Python	Detecting transcription factor footprints and underlying motifs using supervised learning.	+++	Quach and Furey (2017)
DNase2TF	R	Identify footprint candidates from DNase-seq data on user-specified regions.	+	Sung et al. (2014)
HINT-ATAC	Python	Use open chromatin data to identify transcription factor footprints with modifications specific to ATAC-seq data.	+++	Li et al. (2019)
HOMER	Perl; C++	A suite of tools for motif discovery and enrichment.	+++	Heinz et al. (2010)
MEME Suite	Perl; Python	Suite of tools for motif discovery; enrichment; and GO term analyses.	+++	Bailey et al. (2009)
PIQ	Bash; R	Models genome-wide DNase profiles to identify transcription factor binding sites.	++	Sherwood et al. (2014)
TOBIAS	Python	Identify transcription factor footprints.	++	Bentsen et al. (2019)
TRACE	Python	Transcription factor footprinting.	++	Ouyang and Boyle (2019)
Wellington	Python	Identify TF footprints using DNase-seq data.	+++	Piper et al. (2013)

Table 6: **Motif enrichment and transcription factor footprinting tools.**

Another common analysis of differentially accessible regions is *de novo* motif analysis, which

is to look for an overrepresentation of transcription factor motifs in regions of interest relative to some background set. Motif discovery is typically used in analysis of ChIP-seq data, but is also relevant for accessible chromatin peaks with some specificity, such as for a particular cell-type or treatment. Motif discovery has an ongoing field of study for decades, and there are many tools to identify enriched motifs [263–267]. Tools initially designed for ChIP-seq or DNase-seq experiments have been widely applied to ATAC-seq data as well [MEME Suite, 266, HOMER, 267]. There are now dozens or hundreds of individual motif-finding tools [268].

A related approach called *footprinting* explores the microarchitecture of reads *within* peaks to identify physical evidence of bound transcription factors that *decrease* the accessibility at small binding sites (typically under 20 bp) within an overall area of higher accessibility (Table 6) [269]. Following the introduction and rapid adoption of DNase-seq, the number of tools to perform TF footprinting rapidly expanded. A number of these were designed for DNase-seq, but have often been employed using ATAC-seq data successfully [CENTIPEDE, [270]; PIQ, [147]; DNase2TF, [271]; BinDNase, [272]; Wellington, [273]; [274]; TRACE, Ouyang2019]. One advantage of usingtools designed for DNase-seq simply lies in their track record of robustness and widely demonstrated utility, even when applied to ATAC-seq data. Yet, there are unique features of ATAC-seq data including nucleosome positioning information and transposase cleavage biases that can be used to inform on TF footprinting. Newer tools either have specific settings to work with ATAC-seq data or were designed specifically for ATAC-seq and may be more appropriate going forward [DeFCoM, 275, TOBIAS, 276, HINT-ATAC, 277, BiFET, 278].

2.8.7 Nucleosome positioning

Name	Languages	Notes	Docs	Reference
HMMRATAC	Java	Identify nucleosome positioning and leverage ATAC-seq specific read outs to call peaks.	+++	Tarbell and Liu (2019)
NucleoATAC	Python; R	Call nucleosomes using ATAC-seq data.	+++	Schep et al. (2015)
NucTools	Perl; R	Calculate nucleosome occupancy profiles on chromatin accessibility data.	+++	Vainshtein et al. (2017)

Table 7: **Tools to investigate nucleosome positioning.**

Nucleosome positioning is crucial in a number of DNA regulatory processes, particularly gene expression, and may be directly interrogated using ATAC-seq data [55, 279, 280]. ATAC-seq is designed to assay regions of open chromatin; in other words, to identify regions *not currently packaged into nucleosomes*. As a consequence of this, sequenced fragment lengths and alignments occur in structured patterns that inform on the presence and positioning

Name	Languages	Notes	Docs	Reference
Annotatr	R	Annotate summarize and visualize genomic regions.	+++	Cavalcante and Sartor (2017)
BART/BARTweb	Python	Predict factors that bind at cis-regulatory regions.	+++	Wang et al. (2018)
chipenrich	R	Perform gene set enrichment testing using genomic regions.	+++	Welch et al. (2014)
coloc-stats	Python	Perform co-localization analysis of genomic regions.	+++	Simovski et al. (2018)
COLO	JSP	Identify genomic features in close proximity to user-submitted genomic regions.	++	Kim et al. (2015)
FEATnotator	Perl; R	Annotate genomic regions.	++	Podicheti and Mockaitis (2015)
GenomeRunner	.NET	Perform annotation and enrichment of genomic regions against default or custom regulatory regions.	++	Dozmorov et al. (2016)
GenometriCorr	R	Determine spatial correlation between region sets.	++	Favorov et al. (2012)
Genomic Association Tester	Python	Calculate the significance of overlaps between multiple genomic region sets.	+++	Heger et al. (2013)
GIGGLE	C	Genomics search engine to uncover signifcantly shared genomic loci (regions) between data.	+++	Layer et al. (2018)
GLANET	Java; Perl	Genomic loci annotation and enrichment tool between sets of genomic regions.	+++	Otlu et al. (2017)
GREAT	C	Annotate genomic regions.	+++	McLean et al. (2010)
LOLA/LOLAweb	R	Determine significant enrichment between region sets to inform on biological meaning.	+++	Sheffield and Bock (2016)
regioneR	R	Evaluate significant associations between region sets using permutation testing.	+++	Gel et al. (2016)
StereoGene	C++; R	Estimate genome-wide correlation between pairs of genomic features.	++	Stavrovskaya et al. (2017)

Table 8: **Tools to investigate region enrichments**

of nucleosomes (Table 7). Essentially, short ATAC-seq fragments represent nucleosome-free regions, and longer fragments represent nucleosome-associated DNA [170]. The earliest tool [NucleoATAC, 280] reports the position and occupancy of nucleosomes. Building on the fact that this information is inherent to ATAC-seq data, later tools extend the biological information

that can be obtained from a more thorough understanding of nucleosome positioning. The use of nucleosome positioning information may now be easily compared between sample conditions which ultimately allows for concurrent identification of transcription factor binding sites alongside additional epigenetic marks [NucTools, 281]. Furthermore, this information may be leveraged to improve peak calling by incorporating nucleosome positioning and enrichment to more accurately predict true positive open chromatin [HMMRATAC, 258].

2.8.8 Region enrichment

A widely successful analysis type for gene expression data is gene ontology analysis or gene set enrichment analyses, which can be extended to region-based enrichments. In this context, instead of genes as the units of interest, the analysis is done on non-coding regions corresponding to regulatory elements. As chromatin accessibility has increased, so has interest in assigning biological meaning to non-coding loci. Region set enrichment analyses are one approach to this problem. Generally, these tools compare a set of regions of interest (*i.e.*, called peaks) to regions with known biological function. The tools then assess similarity to determine whether there are significant enrichments of overlap between the region sets. This approach can function by identifying significantly enriched GO terms [GREAT, 282], and/or by comparing any previously annotated region set with your unknown peak set [regioneR, 283, LOLA, 284, annotatr, 285, GIGGLE, 286]. Therefore, to assign more meaningful biological relationships to annotated ATAC-seq peaks, one can investigate what specific biological features are correlated or enriched in your peak set (Table 8). These tools and other related tools have been reviewed elsewhere in detail [287, 288].

2.8.9 Single-cell

Although single-cell ATAC-seq (scATAC-seq) is only a few years old [17, 178], the number of available analysis tools has proliferated rapidly (Table 9). A primary challenge to any single-cell sequencing assay is the sparsity of data. For that reason, modifications to general ATAC-seq data processing are necessary. Tools specific to single-cell ATAC-seq analysis include both raw processing pipelines [CellRanger ATAC; BROCKMAN, 289, Scasat, 221, SnapATAC, 290, scATAC-pro, 291] as well as downstream analysis tools, particularly for clustering individual cells into separate cell-type populations [BAP, 179, scABC, 292, SCALE, 293] and identifying transcription factor accessibility [SCRAT, 294, chromVAR, 295, Cicero, 220, cisTopic, 296, scOpen, 297]. Single-cell ATAC-seq analysis is a rapidly changing area, with many of these tools published only within the past year.

Name	Languages	Notes	Docs	Reference
BAP	R; Python	Bead-based scATAC-seq data processing.	++	Lareau et al. (2019)
BROCKMAN	R; Bash; Ruby	Convert genomics data into K-mer words associated with chromatin marks used to compare and identify changes across samples.	++	de Boer and Regev (2018)
Cell Ranger ATAC	NA	Commercial. Set of analysis pipelines for Chromium single cell ATAC-seq.	+++	unpublished
chromVAR	R	Identify transcription factor accessibility in single-cell data. Enables clustering of single-cell ATAC-seq data.	+++	Schep et al. (2017)
Cicero	R	Predict cis-regulatory DNA interactions using single-cell chromatin accessibility data.	+++	Pliner et al. (2018)
cisTopic	R	Identify cell states and cis-regulatory topics from single-cell data.	+++	Bravo Gonzalez-Blas et al.(2019)
scABC	R	Classify single-cell ATAC using unsupervised clustering and identify chromatin regions specific to cell identity.	+	Zamanighomi et al. (2018)
SCALE	Python	Clustering and visualization of single-cell ATAC-seq data into interpretable cell populations.	++	Xiong et al. (2019)
Scasat	Bash; Python; R	Complete pipeline to process scATAC-seq data with simple steps.	+++	Baker et al. (2019)
scATAC-pro	R; Python	Comprehensive pipeline for single cell ATAC-seq analysis.	+++	Yu et al. (2019)
scOpen	Python	Chromatin-accessibility estimation of single-cell ATAC data.	+	Li et al. (2019)
SCRAT	R	Useful for studying single cell heterogeneity. Can identify changes in gene sets or transcription factor binding sites. Includes GUI and web-based service.	+++	Ji et al. (2017)
SnapATAC	R; Python	Single Nucleus Analysis Pipeline for ATAC-seq.	+++	Fang et al.(2019)

Table 9: **Tools for single cell ATAC-seq data processing**

2.9 Conclusion

Chromatin accessibility analysis is becoming increasingly relevant for a range of biological research areas. As scientists realize the richness of chromatin accessibility data, new analytical approaches and tools are being developed. At the same time, chromatin accessibility analysis is now approachable by individuals with a wider range of perspective and experience. This has led to a wide increase in biological results, tools, and analytical approaches.

In our survey of ATAC-seq analysis tools, we identified more than 50 tools employed specifically for ATAC-seq data analysis. In assessing this diverse range of tools, we have found it useful to categorize them by primary aim. Because the diversity and number of available tools and approaches is likely to only increase as ATAC-seq analysis becomes mainstream, we believe it will be important to continue to revisit such tool surveys as the field develops. These summaries provide novices with a basic understanding and starting point, and also give experienced analysts a reference resource to provide ideas for more detailed analysis.

3 PEPPRO: quality control and processing of nascent RNA profiling data (modified from [215])

Following our survey of ATAC-seq analytical tools, we also discovered a dearth of pipelines for processing the emerging field of nascent RNA sequencing. While tools existed to analyze processed data, no unified approach existed to generate the input required for those approaches. Nor were there any standard metrics of nascent RNA-seq quality control or measures of successful nascent sequencing preparations. I was motivated to address these weaknesses by developing a nascent RNA-seq pipeline with novel metrics of sequencing success.

3.1 Background

Steady-state transcription levels are commonly measured by RNA-seq, but there are many advantages to quantifying *nascent* RNA transcripts: First, it measures the transcription process directly, whereas steady-state mRNA levels reflect the balance of mRNA accumulation and turnover. Second, nascent RNA profiling measures not only RNA polymerase occupancy, but also orientation by default, whereas traditional RNA-seq requires specific library preparation steps to capture orientation. Third, nascent RNA profiling measures unstable transcripts, which can be used to infer regulatory element activity and identify promoters and enhancers *de novo* by detecting bidirectional transcription and clustered transcription start sites (TSSs) [210, 298]. Fourth, nascent RNA profiling can be used to determine pausing and RNA polymerase

accumulation within any genomic feature. These advantages have led to growing adoption of global run-on (GRO-seq), precision run-on (PRO-seq), and, most recently, chromatin run-on (ChRO-seq) experiments [202–204]. With increasing data production, we require analysis pipelines for these data types. While tools are available for downstream analysis, such as to identify novel transcriptional units and bidirectionally transcribed regulatory elements [211, 298–302], there is no comprehensive, unified approach to initial sample processing and quality control.

Here, we introduce `PEPPRO`, an analysis pipeline for uniform initial sample processing and novel quality control metrics. `PEPPRO` features include: 1) a serial alignment approach to remove ribosomal DNA reads; 2) nascent transcription-specific quality control outputs; and 3) a modular setup that is easily customizable, allowing modification of individual command settings or even swapping software components by editing human-readable configuration files. `PEPPRO` is compatible with the Portable Encapsulated Projects (PEP) format, which defines a common project metadata description, facilitating interoperability [303]. `PEPPRO` can be easily deployed across multiple samples either locally or via any cluster resource manager, and we also produced a computing environment with all the command-line tools required to run `PEPPRO` using either `docker` or `singularity` with the `bulker` multi-container environment manager [304]. Thus, `PEPPRO` provides a unified, cross-platform pipeline for nascent RNA profiling projects.

A

#	Sample name	Description	Source
1	K562 PRO-seq	high quality PRO-seq data	GSM1480327
2	K562 GRO-seq	high quality GRO-seq data	GSM1480325
3	HelaS3 GRO-seq	low quality across multiple metrics	GSM1558746
4	Jurkat ChRO-seq 1	high quality ChRO-seq data	GSM3309956
5	Jurkat ChRO-seq 2	relatively low complexity	GSM3309957
6	HEK PRO-seq	intact RNA	GSM3618147
7	HEK ARF PRO-seq	degraded RNA	GSM3618143
8	H9 PRO-seq 1		GSM4214080
9	H9 PRO-seq 2		GSM4214081
10	H9 PRO-seq 3	Differential expression analysis	GSM4214082
11	H9 treated PRO-seq 1	observed Δ pause index following treatment	GSM4214083
12	H9 treated PRO-seq 2		GSM4214084
13	H9 treated PRO-seq 3		GSM4214085

B

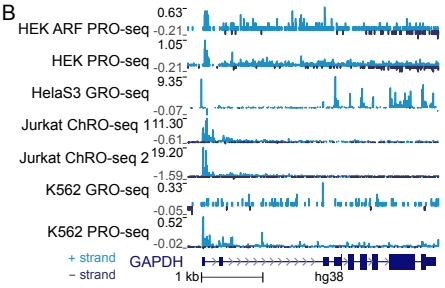

Figure 3.1: **`PEPPRO` test set data table and signal tracks**. A) Table showing the attributes of samples collected for our test set. Complete metadata is available from the `PEPPRO` website. B) Read count normalized signal tracks from published data are visualized within a browser (Scale is per 1M).

3.2 Results

3.2.1 Pipeline overview and data description

`PEPPRO` starts from raw, unaligned reads, and produces a variety of output formats, plots, and quality control metrics. Briefly, pre-alignment steps include removing adapters, deduplicating, trimming, and reverse complementation (Fig. 3.2). `PEPPRO` then uses a serial alignment

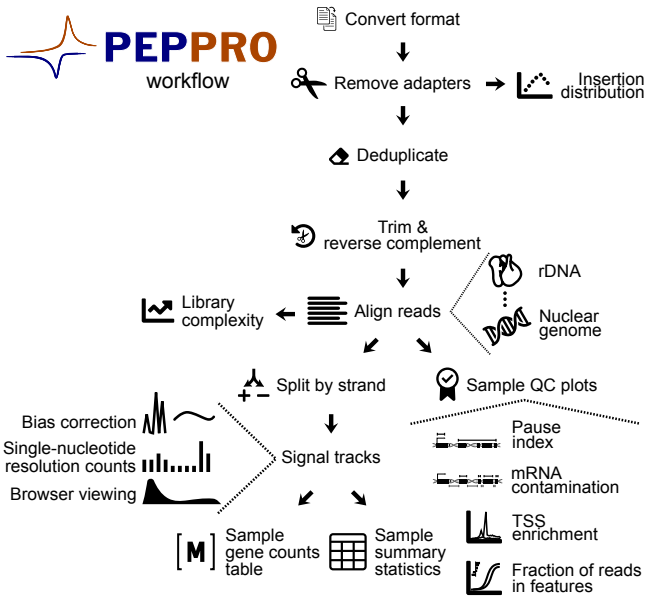

Figure 3.2: **PEPPRO steps for genomic run-on data.** PEPPRO starts from raw sequencing reads and produces a variety of quality control plots and processed output files for more detailed downstream analysis.

strategy to siphon off unwanted reads from rDNA, mtDNA, and any other user-provided decoy sequences. It aligns reads and produces signal intensity tracks as both single-nucleotide counts files and smoothed normalized profiles for visualization. PEPPRO also provides a variety of plots and statistics to assess several aspects of library quality, such as complexity, adapter abundance, RNA integrity and purity, and run-on efficiency (See Methods for complete details).

To evaluate PEPPRO on different library types, we assembled a test set of run-on libraries with diverse characteristics (Fig. 3.1A). Our test set includes 7 previously published libraries: 2 ChRO-seq, 2 GRO-seq, and 3 PRO-seq [204, 305–307]. We ran each of these samples through PEPPRO as a test case and visualized the data in a genome browser (Fig. 3.1B). To demonstrate PEPPRO's setup for differential expression analysis, we also generated paired-end PRO-seq libraries from H9 cell culture samples either naive or treated with romidepsin, a histone deacetylase inhibitor (HDACi). This test set therefore provides a range of qualities, protocols, and issues, providing a good test case for demonstrating the novel quality control features of PEPPRO and how to distinguish high-quality samples.

To demonstrate how PEPPRO responds to mRNA contamination, we also generated a set of 11 samples built from a single PRO-seq library (GSM1480327) that we spiked with increasing amounts of RNA-seq data (GSM765405) (Additional file 1: Fig. 3.10). We ran PEPPRO on our public test set, our differential expression test set, and our spike-in set. Results of PEPPRO can be explored in the PEPPRO HTML-based web report, which displays all of the output

statistics and QC plots (see PEPPRO documentation). Here, we describe each plot and statistic produced by PEPPRO.

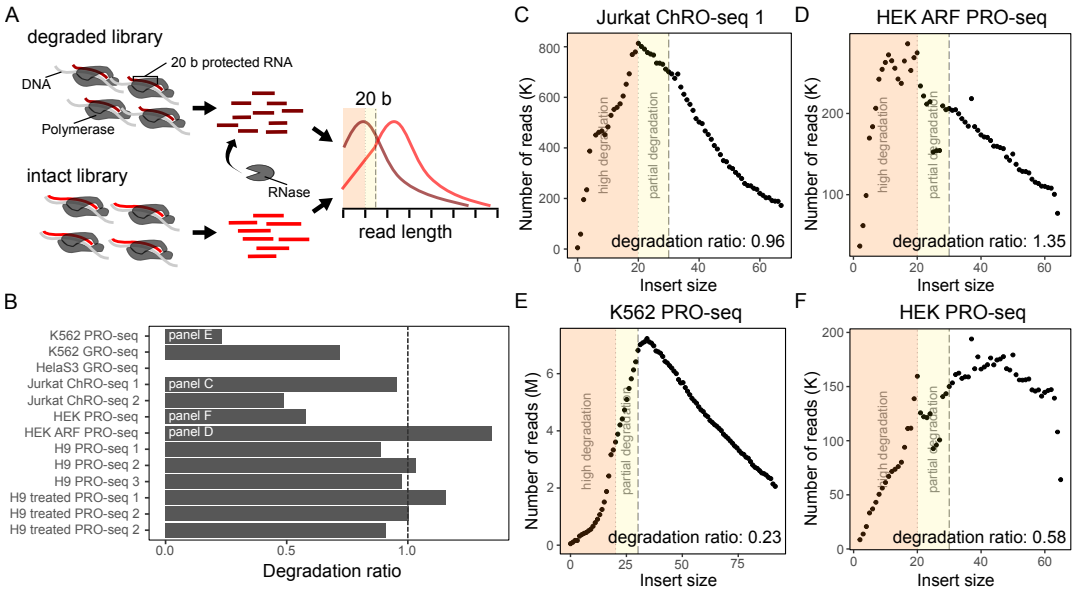

Figure 3.3: **RNA integrity is assessed with degradation ratios and insert sizes.** A) Schematic illustrating intact versus degraded libraries. B) Degradation ratio for test samples (HelaS3 GRO sample could not be calculated; Values less than dashed line (1.0) are considered high quality). C-F) Insert size distributions for: C, a degraded single-end library; D, a degraded paired-end library; E, a non-degraded single-end library; and F, a non-degraded paired-end library (orange shading represents highly degraded reads; yellow shading represents partially degraded reads).

3.2.2 Adapter ratio

A common source of unwanted reads in PRO/GRO/ChRO-seq libraries results from adapter-adapter ligation. These methods require two independent ligation steps to fuse distinct RNA adapters to each end of the nascent RNA molecule. The second ligation can lead to adapter-adapter ligation products that are amplified by PCR. The frequency of adapter-adapter ligation can be reduced by molecular techniques (see Methods), but these are not always possible and many experiments retain adapters in high molar excess, leading to substantial adapter-adapter sequences.

PEPPRO counts and reports the fraction of reads that contain adapter-adapter ligation products, then removes adapter sequences and adapter-adapter ligation sequences before downstream alignment. In our test, all samples had fewer than 50% adapter-adapter ligation reads (Additional file 1: Fig. 3.11). Higher rates do not necessarily reflect lower quality samples, but rather indicate a suboptimal ratio of adapters during the library preparation or exclusion of the gel extraction size selection step. Excess adapters indicate that future sequencing will be less informative, leading to increased depth requirements, and therefore inform on whether to

sequence a library deeper, tweak the adapter ratio in future samples, or include a size selection step. In our hands, we aim for adapter-adapter ligation abundance between 20-50% with no size selection step, or less than 5% if the final library is polyacrylamide gel electrophoresis (PAGE) purified. Libraries with no adapter-adapter ligation indicate that size selection was too stringent, and may actively select against short RNA insertions from specific classes of nascent RNA, such as RNAs from promoter-proximal paused polymerases [308].

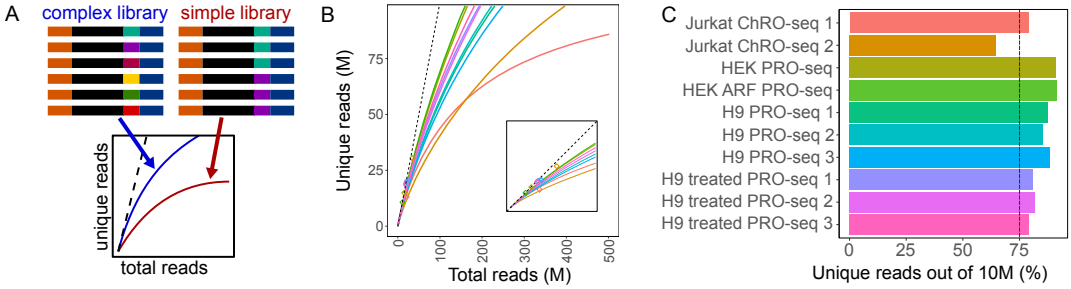

Figure 3.4: **Library complexity is measured with unique read frequency distributions and projections.** A) Schematic demonstrating PCR duplication and library complexity (dashed line represents completely unique library). B) Library complexity traces plot the read count versus externally calculated deduplicated read counts. Deduplication is a prerequisite, so these plots may only be produced for samples with UMIs. Inset zooms to region from 0 to double the maximum number of unique reads. C) The position of curves in panel B at a sequencing depth of 10 million reads (dashed line represents minimum recommended percentage of unique reads).

3.2.3 RNA integrity

A common indicator of RNA sample quality is the level of RNA integrity. RNA integrity can be assessed by plotting the distribution of RNA insert sizes, which will be smaller when RNA is degraded. For a highly degraded library, we expect insert sizes below 20 nucleotides, which corresponds to the length of RNA between the RNA polymerase exit channel and 3′ RNA end. These nucleotides are sterically protected from degradation [309], so high frequency of insert sizes below 20 indicates that degradation occurred after the run-on step [204] (Fig. 3.3A).

`PEPPRO` uses a novel method to calculate the insert size distribution that applies to both single- and paired-end data (see Methods). `PEPPRO` reports the ratio of insert sizes from 10-20 nucleotides versus 30-40 nucleotides, which measures RNA integrity because more degraded libraries have higher frequency of reads of length 10-20, whereas less degraded libraries have more reads of length 30-40. Using our test set, we found that PRO-seq libraries with a ratio < 1 should be considered high quality (Fig. 3.3B). A single-end ChRO-seq library that was intentionally degraded with RNase prior to the run on step [204] has a degradation ratio near 1 with a insertion distribution plot showing a peak at 20 nucleotides (Fig. 3.3C). A poor quality paired-end PRO-seq library contains many RNA species falling within the 10-20

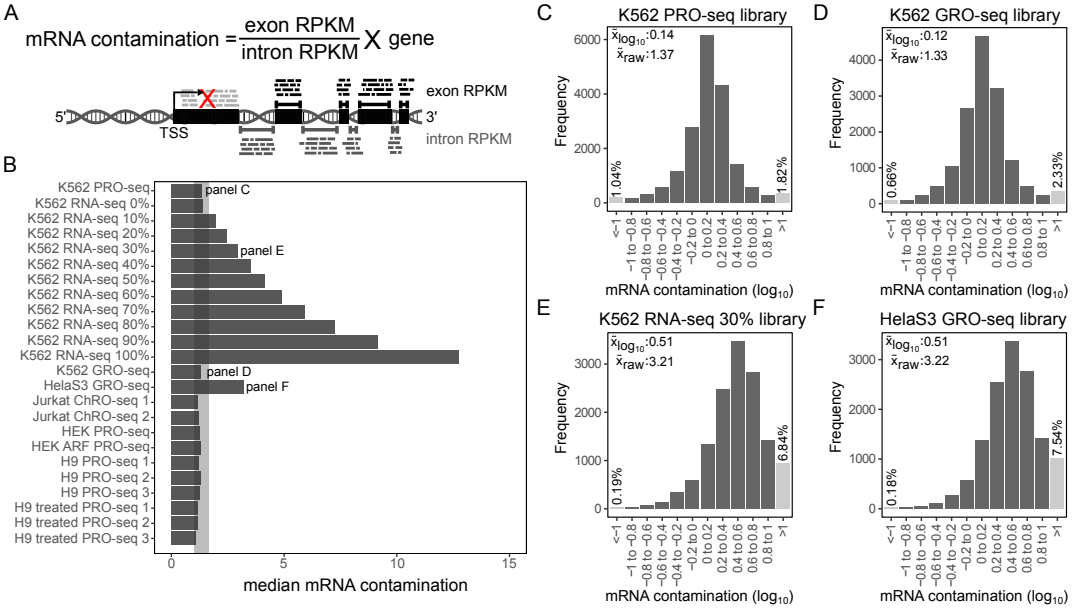

Figure 3.5: **Nascent RNA purity is assessed with the exon-intron ratio**. A) Schematic demonstrating mRNA contamination calculation. X represents the exclusion of the first exon in the calculation. B) Median mRNA contamination metric for test set samples (Shaded region represents recommended range (1-1.8)). C) Histogram showing the distribution of mRNA contamination score across genes in the K562 PRO-seq sample. D) As in panel C for a GRO-seq library. E) mRNA contamination distribution for K562 PRO-seq spiked with 30% K562 RNA-seq. F) mRNA contamination distribution for HelaS3 GRO-seq is comparable to the 30% RNA-seq spike-in sample. $\tilde{x}$ = median; $\bar{x}$= mean)

range (Fig. 3.3D). High-quality libraries show plots that peak outside of the sub-20-nucleotide degradation zone (Fig. 3.3E, F).

3.2.4 Library complexity

Library complexity measures the uniqueness of molecules in a sequencing library (Fig. 3.4A). For conventional RNA-seq, shearing is random, so paired-end reads with the same start and end coordinates may be assumed to be PCR duplicates. In contrast, in PRO-seq, transcription start sites account for many of the 5′ RNA ends, and promoter proximal pause sites can focus the 3′ end of the RNA [203], so independent insertions with the same end points are not necessarily PCR duplicates. As a result, unfortunately, this means it is not possible to calculate complexity generally.

Recent PRO-seq protocols resolve this by incorporating a unique molecular identifier (UMI) into the 3′ adapter, which `PEPPRO` uses to distinguish between PCR duplicates and independent RNA molecules with identical ends. For data that includes UMIs, `PEPPRO` accommodates multiple software packages for read deduplication, including `seqkit` [310] and `fqdedup` [311]. `PEPPRO` calculates library complexity at the current depth, reporting the percentage of PCR duplicates. In our test samples, we found that libraries with at least 75% of reads unique

at a sequencing depth of 10 million can be considered high quality (Fig. 3.4C). PEPPRO also invokes preseq [312] to project the unique fraction of the library if sequenced at higher depth (Fig. 3.4B). These metrics provide a direct measure of library complexity and allow the user to determine value of additional sequencing. However, because nascent RNA reads cannot be effectively deduplicated using the standard approach applied to traditional RNA-seq, complexity metrics are only calculated for samples with UMIs.

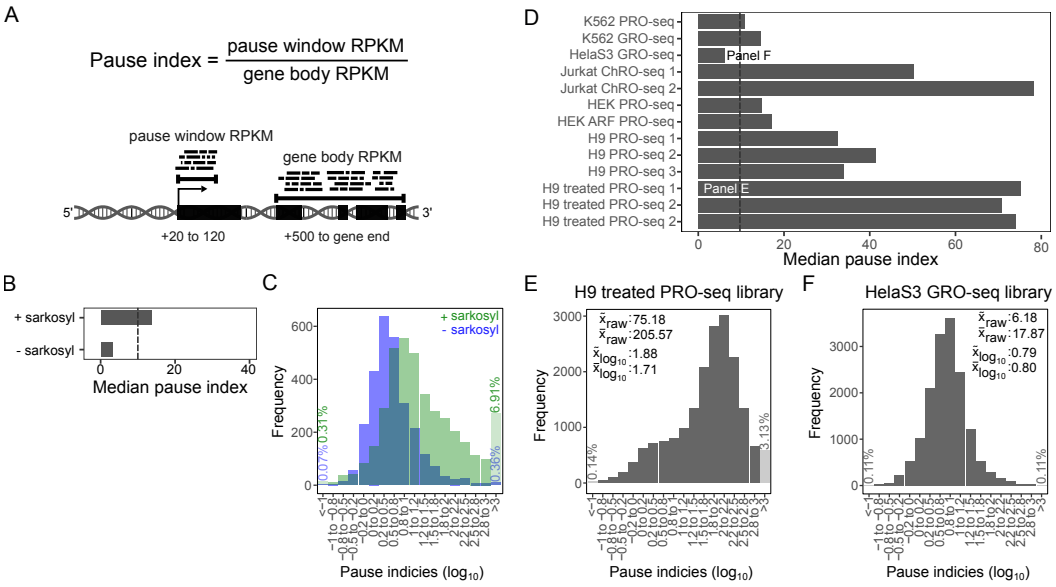

Figure 3.6: **Run-on efficiency is measured with pause indices**. A) Schematic demonstrating pause index calculation. B) Pause index values for *Drosophila melanogaster* GRO-seq libraries with (GSM577247) or without sarkosyl (GSM577248). C) The histogram of pause index values is shifted to the right upon addition of sarkosyl in GRO-seq libraries. D) Pause index values for test set samples (Values above the dashed line are recommended). E) High pause index identified in H9 treated PRO-seq. F) Low pause index from HelaS3 GRO-seq. ($\tilde{x}$ = median; $\bar{x}$=mean)

3.2.5 Nascent RNA purity

One challenge specific to nascent RNA sequencing is ensuring that the library targets nascent RNA specifically, which requires eliminating the more abundant processed rRNA, tRNA, and mRNA transcripts. Early run-on protocols included 3 successive affinity purifications, resulting in 10,000-fold enrichment over mRNA and over 98% purity of nascent RNA [202, 203]. Newer run-on protocols recommend fewer affinity purifications [307]. Therefore, assessing the efficiency of nascent enrichment is a useful quality control output.

To estimate the *nascent purity* of RNA, PEPPRO provides two results: an mRNA contamination metric and a rDNA alignment rate. First, PEPPRO assesses nascent RNA purity by calculating the exon to intron read density ratio (Fig. 3.5A). A nascent RNA sequencing library without polymerase pausing would have a ratio of exon density to intron density of ≈ 1. Because

promoter-proximal pausing inflates this ratio, PEPPRO excludes the first exon from this calculation. In our test samples, the median exon-intron ratio is between 1.0 and 1.8 for high quality libraries (Fig. 3.5B). Our *in silico* spike-in of conventional RNA-seq increases this ratio proportionally to the level of mRNA contamination (Fig. 3.5B). This ratio varies substantially among genes and PEPPRO produces histograms to compare in more detail among samples (Fig. 3.5C-F). By comparing these values to the spike-in experiment, we can estimate the level of mRNA contamination of a library (Fig. 3.5E, F). A second measure of nascent purity is to evaluate relative rRNA abundance.

Since rRNA represents the vast majority of stable RNA species in a cell, overrepresentation of rRNA reads indicates poor nascent RNA enrichment. We find that high-quality nascent RNA libraries typically have less than 20% rRNA alignment (Additional file 1: Fig. 3.12). In contrast, between 70% and 80% of mature RNA in a cell is rRNA. Therefore, the ratio of rRNA aligned reads compared to the all other reads reflects mature RNA contamination. To demonstrate, we calculated the correlation between the exon-intron read density ratio and the rRNA-to-aligned-reads ratio using the primary set of test samples with additional samples (GSE126919) to increase power. We found these two measures are significantly correlated (Additional file 1: Fig. 3.13). Exon-intron read density ratio is a more robust measure of nascent RNA purity, as the fraction of nascent rRNA transcription is likely to be distinct among cell lines. However, PEPPRO still reports the rDNA alignment ratio as an orthogonal measure of nascent RNA purity and overall library quality.

3.2.6 Run-on efficiency

Another quality metric for run-on experiments is run-on efficiency. Typically, gene-body polymerases extend efficiently during the nuclear run-on step, but promoter-proximal paused polymerases require either high salt or detergent to do so [313, 314]. Because these treatments vary, leading to varying run-on efficiency, PEPPRO employs two methods to assess run-on efficiency: *pause index* and *TSS enrichment*. First, we define the pause index as the ratio of the density of reads in the *pausing* region versus the density in the corresponding gene body (Fig. 3.6A; see Methods). PEPPRO plots the frequency distribution of the pause index across genes. A greater pause index indicates a more efficient run-on, as a higher value indicates that paused polymerases efficiently incorporate the modified NTPs. As test of this metric, we analyzed GRO-seq data that was generated in the presence and absence of the anionic detergent Sarkysol [314]. Paused polymerases necessitate detergent to run on and incorporate NTPs efficiently, thus the pause index drops substantially in the absence of Sarkysol (Fig. 3.6B,C). We found in our test samples that an efficient run-on process has a median pause

index greater than 10 (Fig. 3.6D). For more detail, PEPPRO produces frequency distribution plots that show an exponential distribution among genes for an efficient library (or a normal distribution on a log scale, Fig. 3.6E) and a shifted distribution for an inefficient run-on (Fig. 3.6F).

As a second assessment of run-on efficiency, PEPPRO aggregates sequencing reads at TSSs to plot and calculate a TSS enrichment score. PEPPRO plots aggregated reads 2000 bases upstream and downstream of a reference set of TSSs. The normalized TSS enrichment score is calculated by taking the average base coverage in a 100 bp window around the peak divided by the average coverage in the first 200 bases. Efficient TSS plots show a characteristic PRO-seq pattern with an upstream peak for divergently transcribing polymerases and a prominent peak representing canonical paused polymerases (Additional file 1: Fig. 3.14). PEPPRO also summarizes these values across samples.

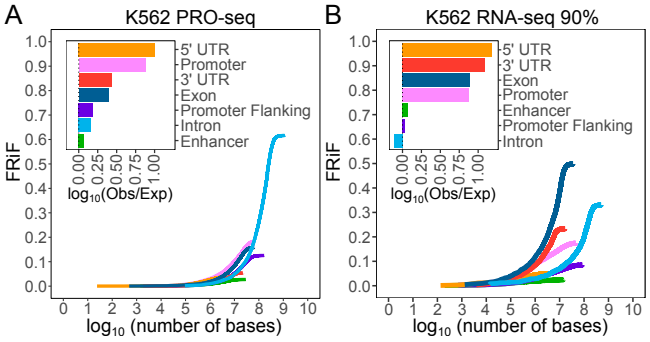

Figure 3.7: **Fraction of reads in genomic features**. A) K562 PRO-seq represents a "good" cumulative fraction of reads in features (cFRiF) and fraction of reads in features (FRiF) plot. B) K562 PRO-seq with 90% K562 RNA-seq spike-in represents a "bad" FRiF/PRiF.

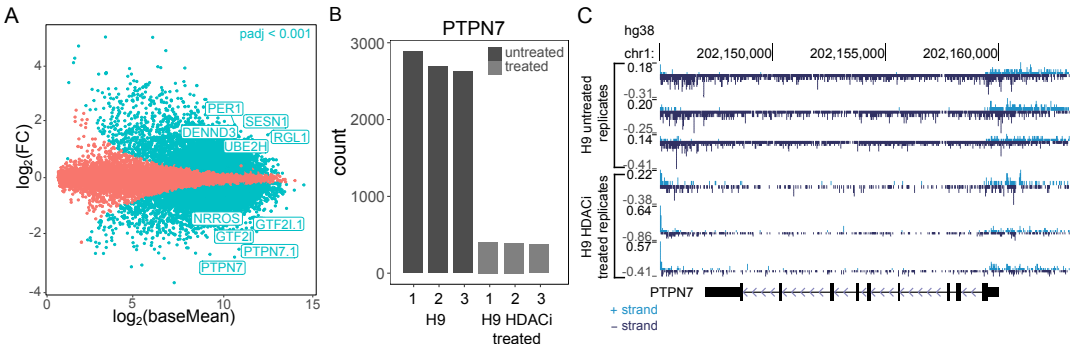

Figure 3.8: **Differential analysis with the PEPPRO counts matrix**. A) MA plot between H9 DMSO versus H9 200nM romidepsin treated PRO-seq libaries (dots = genes; top 10 most significant genes labeled; n=3/treatment). B) Most significantly differential gene count differences. C) Read count normalized signal tracks from the differential analysis (Scale is per 1M).

3.2.7 Read feature distributions

PEPPRO also produces plots to visualize the *fraction of reads in features*, or FRiF. The cumulative FRiF (cFRiF) plot provides an information-dense look into the genomic distribution of reads relative to genomic features. This analysis is a generalization of the more common *fraction of reads in peaks* (FRiP) plots produced for other data types [315] with two key differences: First, it shows how the reads are distributed among different features, not just peaks; and second, it uses a cumulative distribution to visualize how quickly the final read count is accumulated in features of a given type. To calculate the FRiF, PEPPRO overlaps each read with a feature set of genomic annotations, including: enhancers, promoters, promoter flanking regions, 5′ UTR, 3′ UTR, exons, and introns (Fig. 3.7). The individual feature elements are then sorted by read count, and for each feature, we traverse the sorted list and calculate the cumulative sum of reads found in that feature divided by the total number of aligned reads. We plot the read fraction against the log_{10} transformed cumulative size of all loci for each feature. This allows the identification of features that are enriched for reads with fewer total features and total genomic space. Additionally, PEPPRO calculates the non-cumulative FRiF by taking the log_{10} of the number of observed bases covered in each feature over the number of expected bases in each feature to identify enriched genomic features (Fig. 3.7).

In our test samples, high-quality libraries have a characteristic pattern with slow accumulation but high total of reads in introns, and fast accumulation but lower total of reads in promoter elements. ChRO-seq libraries have an increased promoter emphasis and higher mRNA contamination indicated by an increase in reads in promoters and exons at the cost of reads in introns and promoter flanking regions (Additional file 1: Fig. 3.15). Additionally, the RNA-seq spike-in samples demonstrate the increasing prevalence of exonic reads and 3′ UTR at the cost of intronic sequences (Additional file 1: Fig. 3.16). These plots are therefore a useful general-purpose quality control tool that reveal substantial information about a sample in a concise visualization.

3.2.8 Differential expression

The focus of PEPPRO is in the pre-processing relevant for any type of biological project. The output of PEPPRO sets the stage for downstream analysis specific to a particular biological question. Perhaps the most common downstream application of nascent transcription data is differential expression analysis. PEPPRO allows the user to easily run a differential comparison using dedicated software like the DESeq2 bioconductor package [261]. To demonstrate this,

we included PRO-seq libraries from H9 human cutaneous T-cell lymphoma cell lines treated with either DMSO (n=3) or an HDAC inhibitor (n=3).

To facilitate differential expression analysis, `PEPPRO` produces a project-level counts table that may be loaded in R using `pepr`, and, in a few lines of code, converted quickly into DEseq data sets ready for downstream DESeq analyses (See Additional file 1: R code to generate a gene counts table). Using this approach, we ran a differential expression analysis comparing romidepsin-treated against untreated samples (Fig. 3.8A). We identified many genes with significantly different read coverage. As an example, the PTPN7 gene showed clear differences in counts (Fig. 3.8B), which we can further visualize using the browser track outputs generated by `PEPPRO` (Fig. 3.8C). This analysis demonstrates how simple it is to ask a downstream biological question starting from the output produced by `PEPPRO`.

Metric	Recommended value
Degradation ratio	< 1
rDNA alignment rate	< 20%
Pause index	> 10
mRNA contamination	1 - 1.8
% uniformative adapter reads (PAGE)	< 5%
% uniformative adapter reads (w/o PAGE)	20 - 50%
% unique at 10M reads	> 75%

Figure 3.9: **Recommendation table.** Based on our experience processing both high- and low-quality nascent RNA libraries, these are our recommended values for high-quality PRO-seq libraries.

3.2.9 Metric robustness

To evaluate the robustness of our metrics across sequencing depth and library complexity, we ran `PEPPRO` on subsampled single-end and paired-end with UMI PRO-seq libraries (Additional file 1: Fig. 3.17, Fig. 3.18). Our metrics remained constant across sequencing depth from as few as 10M reads to well over 100M (Additional file 1: Fig. 3.19, Fig. 3.20). We also generated synthetic low complexity paired-end with UMI PRO-seq libraries and our metrics remain robust to reductions in library complexity (Additional file 1: Fig. 3.21).

Because our metrics are based on specific source annotation files, we also investigated the effect of alternative annotation source files. To illustrate, we recalculated exon:intron density ratios and pause indicies using UCSC RefSeq, Ensembl, and GENCODE gene set annotation files. While specific values per sample may have minor changes, as would be expected, the relationship between samples is consistent (Additional file 1: Fig. 3.22).

3.3 Conclusions

PEPPRO is an efficient, user-friendly PRO/GRO/ChRO-seq pipeline that produces novel, integral quality control plots and signal tracks that provide a comprehensive starting point for further downstream analysis. The included quality control metrics inform on library complexity, RNA integrity, nascent RNA purity, and run-on efficiency with theoretical and empirical recommended values (Fig. 3.9). PEPPRO is uniquely flexible, allowing pipeline users to serially align to multiple genomes, to select from multiple bioinformatic tools, and providing a convenient configurable interface so a user can adjust parameters for individual pipeline tasks. Furthermore, PEPPRO reads projects in PEP format, a standardized, well-described project definition format, providing an interface with Python and R APIs to simplify downstream analysis.

PEPPRO is easily deployable on any compute infrastructure, from a laptop to a compute cluster. It is thereby inherently expandable from single to multi-sample analyses with both group level and individual sample level quality control reporting. By design, PEPPRO enables simple restarts at any step in the process should the pipeline be interrupted. At multiple steps within the pipeline, different software options exist creating a swappable pipeline flow path with individual steps adaptable to future changes in the field. PEPPRO is a rapid, flexible, and portable PRO/GRO/ChRO-seq project analysis pipeline providing a standardized foundation for more advanced inquiries.

3.4 Availability of data and materials

Documentation on the Portable Encapsulated Project (PEP) standard may be found at pep.databio.org. Refgenie documentation and pre-built reference genomes are available at refgenie.databio.org. The PEPPRO documentation, including links to an HTML report for the test samples, is hosted at peppro.databio.org, and source code is available at github.com/databio/peppro and archived under DOI 10.5281/zenodo.4542304 [316].

Primary analyses data were downloaded from GEO accession numbers GSM1480327 [305], GSM1480325 [305], GSM1558746 [306], GSM3309956, GSM3309957 [204], GSM3618147, GSM3618143 [307], GSM4214080, GSM4214081, GSM4214082, GSM4214083, GSM4214084, GSM4214085 [317]. The sarkosyl analysis used data downloaded from GEO accession numbers GSM577247 and GSM577248 [314]. Data for the RNA-seq spike-in analysis was downloaded from GEO accession numbers GSM1480327 [305] and GSM765405 [318]. Additional data for the rDNA to mRNA contamination correlation analysis was downloaded from GEO accession GSE126919 [307].

3.5 Methods

3.5.1 Pipeline implementation

The PEPPRO pipeline is a python script (peppro.py) runnable from the command-line. PEPPRO provides restartability, file integrity protection, logging, monitoring, and other features. Individual pipeline settings can be configured using a pipeline configuration file (peppro.yaml), which enables a user to specify absolute or relative paths to installed software and parameterize alignment and filtering software tools. Required software includes several Python packages (cutadapt[319], looper, numpy[320], pandas[321], pararead, pypiper, and refgenie[322]) and R packages (installed via the included PEPPROr R package) in addition to some common bioinformatics tools including bedtools [323], bigWigCat [324], bowtie2 [325], fastq-pair [326], flash [327], picard, preseq [312], seqkit [310], samtools [328], seqtk, and wigToBigWig [324]. This configuration file will work out-of-the-box for research environments that include required software in the shell PATH, but may be configured to fit any computing environment and is adaptable to project-specific parameterization needs.

3.5.2 Refgenie reference assembly resources

Several PEPPRO steps require generic reference genome assembly files, such as sequence indexes and annotation files. For example, alignment with bowtie2 requires bowtie2 indexes, and feature annotation to calculate fraction of reads in features requires a feature annotation. To simplify and standardize these assembly resources, PEPPRO uses *refgenie*. Refgenie is a reference genome assembly asset manager that streamlines downloading, building, and using data files related to reference genomes [322]. Refgenie includes recipes for building genome indexes and genome assets as well as downloads of pre-indexed genomes and assets for common assemblies. Refgenie enables easy generation of new standard reference genomes as needed. For a complete analysis, PEPPRO requires a number of refgenie managed assets. Those assets as defined by refgenie are: fasta, bowtie2_index, ensembl_gtf, ensembl_rb, refgene_anno, and feat_annotation. If building these assets manually, they separately require a genome fasta file, a gene set annotation file from RefGene, an Ensembl gene set annotation file in GTF format, and an Ensembl regulatory build annotation file. Finally, using PEPPRO with seqOutBias requires the additional refgenie tallymer_index asset of the same read length as the data.

3.5.3 Adapter-adapter ligation product abundance

Adapter-adapter ligation products show up in run-on libraries because there are two independent ligation steps. Sequencing these products is uninformative, and so there are several molecular approaches used to reduce their abundance in a sequencing library. All protocols include an inverted dT on the 3′ end of the 3′ adapter, and also do not phosphorylate the 5′ end of the 5′ adapter. Many protocols include a size-selection gel extraction step to purify the library from a prominent adapter-adapter ligation species.

`PEPPRO` calculates adapter-adapter ligation products directly from `cutadapt` output, and the default `-m` value for this step is the length of the UMI plus two nucleotides. Therefore, if RNA insertions fewer than three nucleotides in length are present in the library, these are treated as adapter-adapter ligation products.

3.5.4 RNA insert size distribution and degradation

For both single and paired end data, the RNA insert size distribution is calculated prior to alignment. For single end data, the calculation is derived only from sequences that contain adapter sequence, which is output directly from `cutadapt` [319]. `PEPPRO` plots the inverse `cutadapt` report fragment lengths against the `cutadapt` fragment counts. If there is a known UMI, based on user input, that length is subtracted from reported `cutadapt` fragment lengths. As a consequence of this distribution, we can establish a measure of library integrity by evaluating the sum of fragments between 10-20 bases versus the sum of fragments between 30-40 bases in length. The higher this degradation ratio, the more degraded the library.

Paired end sequencing files often have shorter reads because a standard 75 base sequencing cartridge can be used for two paired end reads that are each 38 nucleotides in length. Therefore, many fewer of the reads derived from either end of the molecule extend into the adapter sequence. To address this issue, we incorporate a step that fuses overlapping reads using `flash`[327]. Therefore, if two paired end reads contain overlapping sequence, the reads are combined and the insert size is calculated directly from the fused reads and output directly from `flash`. This distribution is plotted identically to the single end reads and degradation is calculated in the same manner. This degradation ratio metric is uniform between single-end or paired-end libraries and is reported prior to any alignment steps, minimizing influences from extensive file processing or alignment eccentricities.

3.5.5 Excluding size selection skews metrics

Recent PRO-seq protocols, including the H9 libraries we generated, exclude the PAGE size selection step that removes adapter-adapter ligation products [307]. Size selection can potentially bias against small RNA insertions. The previous two metrics: adapter-adapter abundance and degradation ratio are naturally skewed toward the undesirable range if libraries are constructed without size selection. Adapter abundance is skewed because the sole purpose of size selection is to remove the adapter species, but these uninformative reads are of minimal concern and can be overcome by increasing sequencing depth. Degradation ratio is skewed higher because the size selection is not perfect and insert sizes in the range of 10-20 are preferentially selected against relative to those in the 30-40 range. Therefore, while we provide recommendations for optimal degradation ratios, this metric is not necessarily comparable between library preparation protocols and a higher ratio is expected for protocols that exclude size selection.

3.5.6 Removing UMI and reverse complementation

In a typical sequencing library, low library complexity is indicated by high levels of PCR duplicates. Conventional methods remove independent paired-end reads that map to the same genomic positions. This method works reasonably well for molecular genomics data sets with random nucleic acid cleavage. However, in PRO-seq, transcription start sites account for many of the 5′ RNA ends and polymerases pause downstream in a focused region [203]. Consequently, independent insertions with the same end points are common, especially in the promoter-proximal region. To solve this, PRO-seq protocols incorporate a unique molecular identifier (UMI) into the 3′ adapter to distinguish between PCR duplicates and independent insertions with shared ends. `PEPPRO` removes PCR duplicates only if UMIs are provided.

Following the removal of PCR duplicates, the UMI is trimmed. For run-on experiments where the sequencing primer sequences the 3′ end of the original RNA molecule, reverse complementation is performed. As only the first read contains a UMI in paired-end experiments, the second reads skip UMI trimming. Both steps are performed using either `seqtk` (https://github.com/lh3/seqtk) or `fastx` (https://github.com/agordon/fastx_toolkit), depending on user preference. Because reads are processed uniquely for first and second reads in a paired-end experiment, reads must be re-paired prior to alignment. `PEPPRO` uses the optimized implementation `fastq-pair` [326] to re-pair desynchronized read files.

3.5.7 Serial alignments

Following re-pairing, or starting from processed single-end reads, `PEPPRO` performs a series of preliminary, serial alignments (prealignments) before aligning to the primary reference using `bowtie2` [325]. As a significant portion of nascent transcription includes rDNA, `PEPPRO` defaults to initially aligning all reads to the human rDNA sequence. Not only does this remove rDNA reads from downstream analysis, it improves computational efficiency by aligning the largest read pool to a small genome and reduces that read pool for subsequent steps. The user can specify any number of additional genomes to align to prior to primary alignment, which may be used for species contamination, dual-species experiments, repeat model alignments, decoy contamination, or spike-in controls. For serial alignments, `bowtie2` is run with the following parameters `-k 1 -D 20 -R 3 -N 1 -L 20 -i S,1,0.50`, where we are interested primarily in quickly identifying and removing any reads that have a valid alignment to the serial alignment genome (-k 1 parameter). These settings are easily adjusted in the pipeline configuration file (`peppro.yaml`).

Subsequent to these serial alignments, remaining reads are aligned to the primary genome. Primary genome alignment uses the `bowtie2 --very-sensitive` option by default and sets the maximum paired-end fragment length to 2000. The goal with primary alignment is to identify the best valid alignment for reads, sacrificing speed for accuracy. Following primary alignment, low-quality reads are removed using `samtools view -q 10`. As with the initial prealignments, these parameters can be customized by the user in the pipeline configuration file (`peppro.yaml`). Alignment statistics (number of aligned reads and alignment rate) for all serial alignments and primary alignments are reported. For the primary alignment, `PEPPRO` also reports the number of mapped reads, the number removed for quality control, the total efficiency of alignment (aligned reads out of total raw reads), and the read depth. Prior to further downstream analysis, paired-end reads are split into separate read alignment files and only the first read is retained for downstream processing. For both paired-end and single-end experiments, this aligned read file is split by strand with both plus and minus strand aligned files further processed.

3.5.8 Processed signal tracks

Following read processing, alignment, strand separation, and quality control reporting, aligned reads are efficiently converted into strand-specific bigWig files by default. For PRO-seq and similar protocols, reads are reported from the 3′ end and may optionally be scaled by total reads. `PEPPRO` may alternatively use `seqOutBias` [311] to correct enzymatic sequence bias.

Bias is corrected by taking the ratio of genome-wide observed read counts to the expected sequence based counts for each k-mer [311]. K-mer counts take into account mappability at a given read length using Genome Tools' `Tallymer` program [329]. Correcting for enzymatic bias can be important as bias from T4 RNA Ligase used in PRO-seq protocols can yield erroneous conclusions [311]. As such, we recommend using `seqOutBias` for bias correction when analyzing a typical PRO-seq library. Bias correction is especially important when plotting composite profiles over sequence features. Strand specific bigWigs may be visually analyzed using genomic visualization tools and provide a unified starting point for downstream analyses. For example, output bigWig files can be directly loaded into `dREG` to identify regulatory elements defined by bidirectional transcription [298].

3.5.9 Exon-intron ratio plots

`PEPPRO` provides an mRNA contamination histogram for quick visual quality control, and a BED format file containing gene by gene exon:intron ratios for detailed analysis. To calculate this metric, `PEPPRO` utilizes annotation files derived from UCSC RefSeq gene files. Because promoter-proximal pausing inflates these ratios, `PEPPRO` excludes the first exon from the calculation. Otherwise, the reads per kilobase per million mapped reads (RPKM) is calculated for all exonic and intronic sequences on a gene by gene basis. Then, the ratio of exon RPKM to intron RPKM is determined *for every gene.* The overall measure, the mRNA contamination metric, is the median of all genic exon:intron density ratios.

3.5.10 Pause index

Pause indices are calculated as the ratio of read density in the promoter proximal region versus read density in the gene body. To calculate these values, `PEPPRO` utilizes annotation files derived from Ensembl gene set files. Pause indices can vary widely depending on the defined pause window and how a pause window is determined (i.e. relative to a TSS or the most dense window proximal to a TSS). `PEPPRO` defines the density within the pause region as the single, most dense window +20-120 bp taken from all annotated TSS isoforms per gene. This is necessary as some genes contain multiple exon 1 annotations and because this region is where most polymerase pausing occurs, `PEPPRO` identifies the predominant exon 1, based on density, and calculates the pause index using this window density. This means that for genes with multiple TSSs, we define the pause window as the region +20-120 bases from *each* identified TSS per gene. We determine the read density at every annotated pause window per gene, and identify the predominant, singular pause window as the pause window with the greatest density. This singular pause window is used to calculate the pause index for that gene.

The corresponding gene body is defined as the region beginning 500 bp downstream from the predominant TSS to the gene end. We found that lowly expressed genes represent a significant portion of genes with a low pause index. At low sequencing depth, these lowly expressed genes experience greater dropout and fluctuation in pause index calculation, skewing the metric upwards at low depth. To address this, we restrict the pause index calculation to the upper 50th percentile of genes by expression, which eliminates the variability due to depth. Finally, `PEPPRO` plots the distribution of pause indices for each remaining gene in a histogram and provides a BED-formatted file containing each gene's pause index for more detailed analyses.

3.5.11 PRO-seq experiments

H9 PRO-seq experiments were conducted as described previously [307]. The HDACi-treated samples were incubated with 200nM romidepsin for 60 minutes prior to harvesting. The control "untreated" samples were treated with DMSO for 60 minutes. We have included these samples as a test to demonstrate differential expression analysis using `PEPPRO`. They also provide additional example libraries for the metrics in general, and unexpectedly, show significant differences in pause index upon treatment.

3.5.12 Synthetic experiments

Synthetic sequencing depth variant libraries were constructed for single-end and paired-end PRO-seq libraries using either the K562 PRO-seq (GSM1480327) or H9 PRO-seq 2 (GSM4214081) as source libraries, respectively. For K562 PRO-seq subsamples, `seqtk sample -s99` was called on the raw fastq files to generate libraries between 2-10%, in 2 percent increments, and between 10-100%, in 10 percent increments. For the H9 PRO-seq libraries, `seqtk sample -s99` was called on the raw fastq files to produce libraries between 10-100%, in 10 percent increments. Lower percentage K562 PRO-seq libraries were generated to yield libraries of total size comparable to low percentage H9 PRO-seq libraries.

RNA-seq spike-in libraries were also produced using the command `seqtk sample -s99` on raw fastq files using combinations of the K562 PRO-seq library utilized prior and a corresponding K562 RNA-seq library (GSM765405). RNA-seq libraries were sampled between 10-100%, in 10 percent increments, and concatenated with the sampled K562 PRO-seq libraries to generate mixed libraries composed of 0-100% RNA-seq.

Low complexity libraries were similarly constructed. Thirty million total read libraries were generated by using `seqtk sample -s99` on the H9 PRO-seq 2 library and sampling at 50, 80, 90, 92, 94, 96, 98, and 100%. At each percentage of original H9 PRO-seq 2 library sample, the

remainder represents duplicates of the original raw reads composing the opposite percentage, producing libraries with varying levels of duplicated reads.

3.6 Supplemental

3.6.1 R code to generate a gene counts table

PEPPRO provides a project level counts table that simplifies downstream analyses. Here, we import the PEPPRO project counts table and construct a DESeq data set in a few lines of code.

```
# 1. Load the PEPPRO R package.
library(PEPPROr)
# 2. Load the PEP project using the project configuration file.
prj = Project("peppro_paper.yaml")
# 3. Load the project gene counts table.
counts = read.csv(file.path(paste0(config(prj)$metadata$output_dir,
                                   "/summary/PEPPRO_countData.csv")))
# 4. Only keep the H9 untreated or H9 HDAC inhibitor treated samples.
counts = counts[,c("geneName", "H9_PRO-seq_1", "H9_PRO-seq_2", "H9_PRO-seq_3",
                   "H9_treated_PRO-seq_1", "H9_treated_PRO-seq_2",
                   "H9_treated_PRO-seq_3")]
# 5. Convert the counts table to a matrix by removing the gene name column.
count_matrix = as.matrix(counts[,-"geneName"])
# 6. Set the rownames of the matrix object to be the gene names.
rownames(count_matrix) = counts$geneName
# 7. Create a data.frame that defines the sample information.
coldata = data.frame(condition=c(rep("untreated", 3), rep("treated", 3)))
# 8. Set the rownames of the sample information data.frame to match the counts matrix.
rownames(coldata) = colnames(count_matrix)
# 9. Load the DESeq2 package.
library("DESeq2")
# 10. Create a DESeq data set from our counts matrix and the sample information data.frame
dds = DESeqDataSetFromMatrix(countData = count_matrix,
                             colData = coldata,
                             design = \~ condition)
```

3.6.2 Supplemental figures

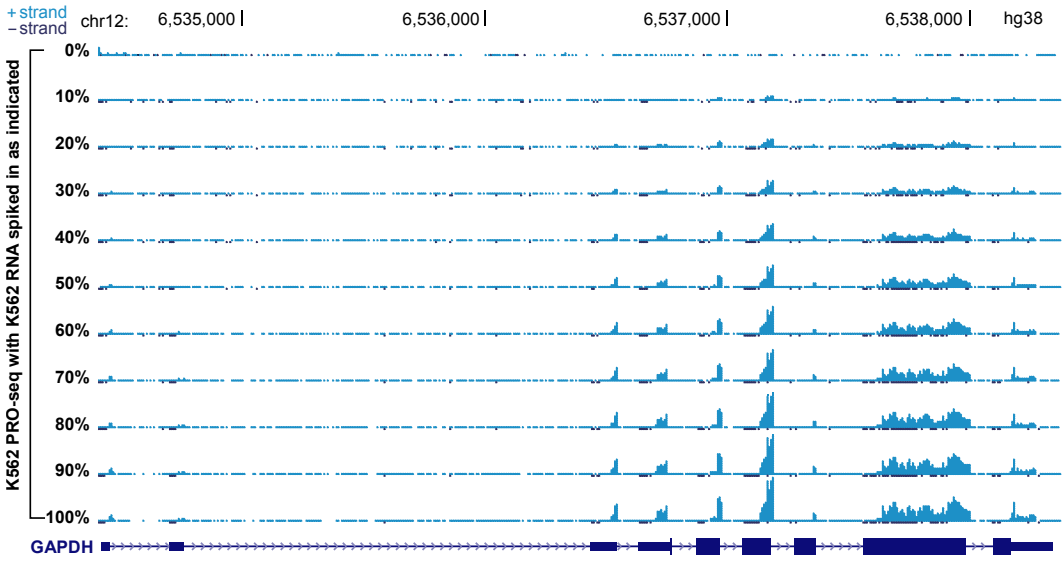

Figure 3.10: **K562 RNA-seq spike-in signal tracks show increasing exonic coverage**. GAPDH exonic coverage is enriched as the percentage of RNA-seq reads increases, and is visualized particularly well at exons 6 and 8. Each sample library is composed of 70M total reads. Scale for each track is 1000 to -20.

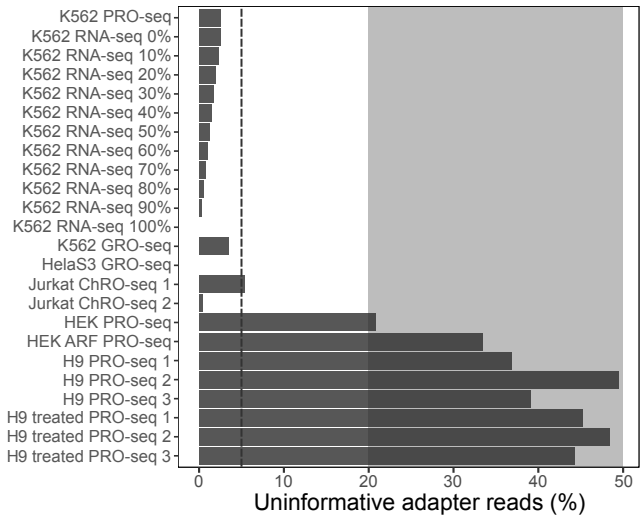

Figure 3.11: **Percentage of uninformative adapter reads following adapter removal for test set samples**. The HEK and H9 libraries contain more adapter-adapter reads because PAGE-mediated size selection was excluded from the protocol (Values below the dashed line are generally recommended for PAGE-purified libraries. Shaded region represents the recommended abundance of adapter reads for libraries *without* PAGE purification.)

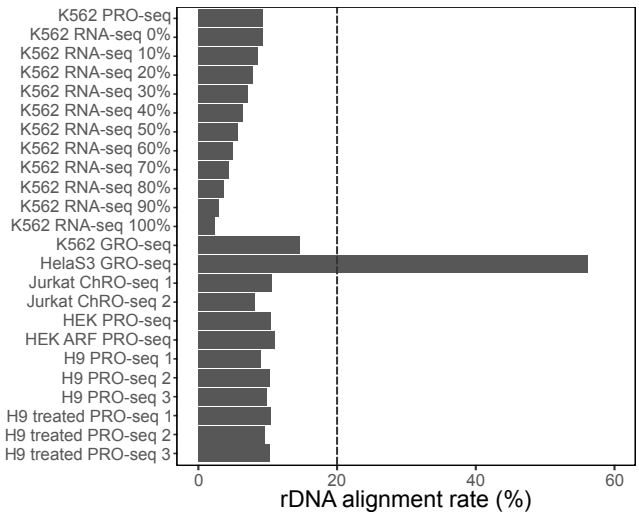

Figure 3.12: **Ribosomal DNA alignment rates for test set samples**. The HelaS3 GRO-seq sample is highly enriched for ribosomal RNA transcripts compared to other test samples. Values below the dashed line are recommended.

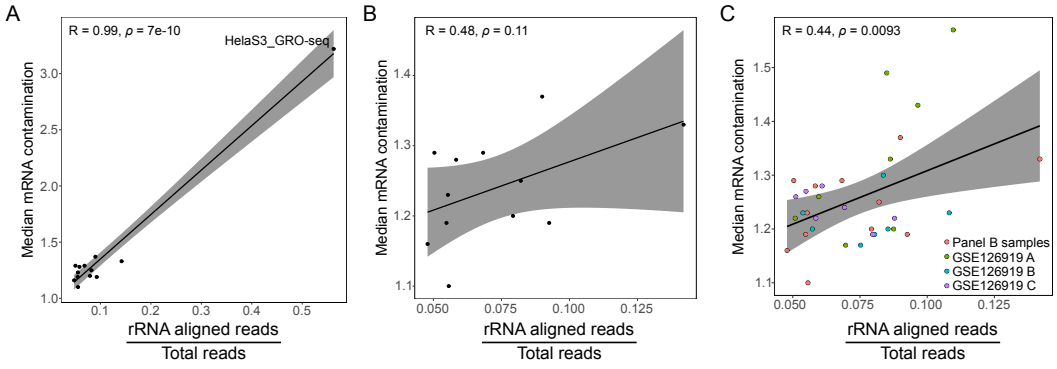

Figure 3.13: **Abundance of rDNA to total reads is correlated with mature RNA contamination.** Correlation plot between the measure of mRNA contamination (median exon:intron density) and the ratio of rDNA aligned reads to total reads for: A) all primary samples (*excludes RNA-seq spike-in experiment due to ribosomal depletion inherent in RNA-seq library preparation), B) primary samples excluding the known outlier HelaS3_GRO-seq sample, C) all samples in panel B *and* all non-redundant samples from GSE126919 including three cellular subclones (A, B, and C) to demonstrate possible differences due to cell lines. Test for association determined with Pearson's product moment correlation coefficient.

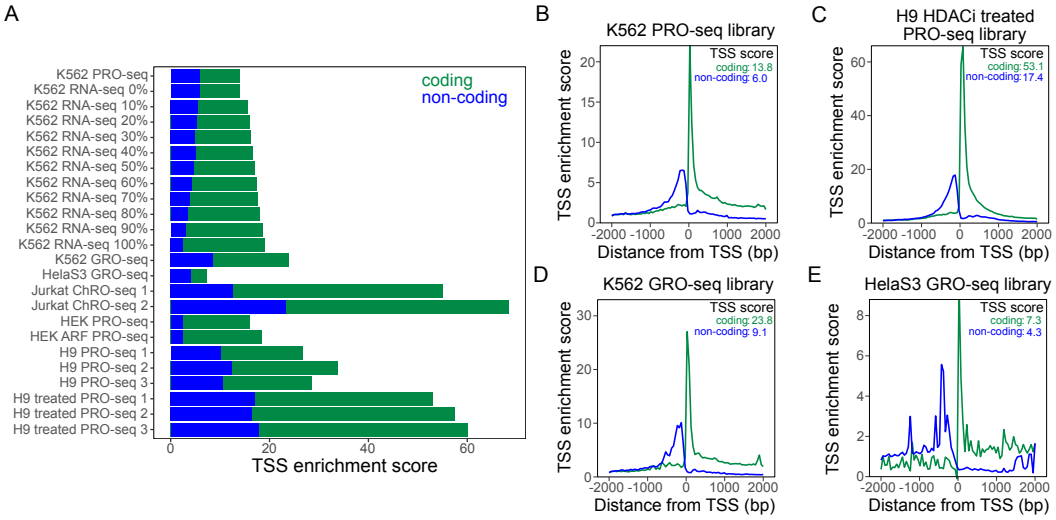

Figure 3.14: **TSS enrichment**. A) TSS enrichment scores for test set samples. B) Representative high-quality PRO-seq TSS enrichment plot. C) TSS enrichment plot in romidepsin treated PRO-seq library. D) Representative high quality GRO-seq TSS enrichment plot E) Representative example of lower quality GRO-seq TSS enrichment plot.

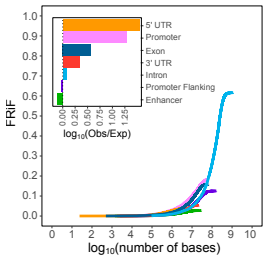

Figure 3.15: **Fraction of Reads in Features in ChRO-seq**. Cumulative FRiF and FRiF (inset) plots for example Jurkat ChRO-seq 1 library test sample shows increased enrichment of promoter sequences.

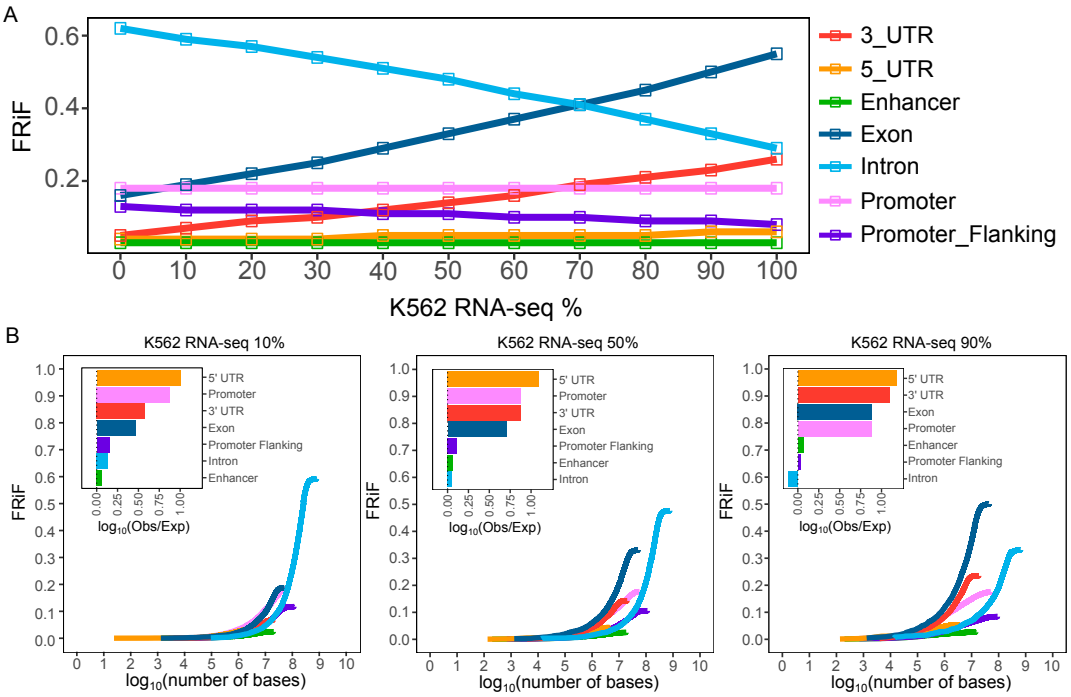

Figure 3.16: **RNA-seq spike-in shows how FRiF changes with mRNA contamination**. A) Increasing percentages of RNA-seq spike-in lead to changes in the fraction of reads in features (FRiF). B) Cumulative FRiF plots at 10%, 50%, and 90% RNA-seq spike-in. Plot insets represent the expected versus observed fraction of reads in genomic features. Each spike-in library contains 70 million total reads.

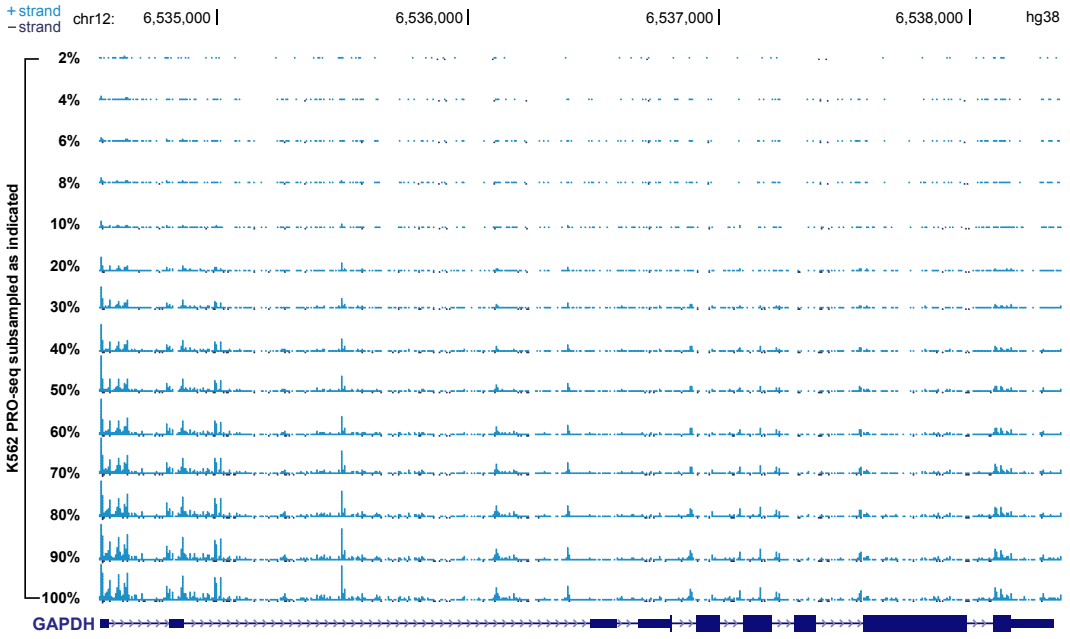

Figure 3.17: **K562 PRO-seq signal tracks show increasing coverage with depth**. Incrementally subsampled K562 PRO-seq library signal tracks display reduced relative coverage at a representative locus (GAPDH). Fixed scale for each track is 100 to -10.

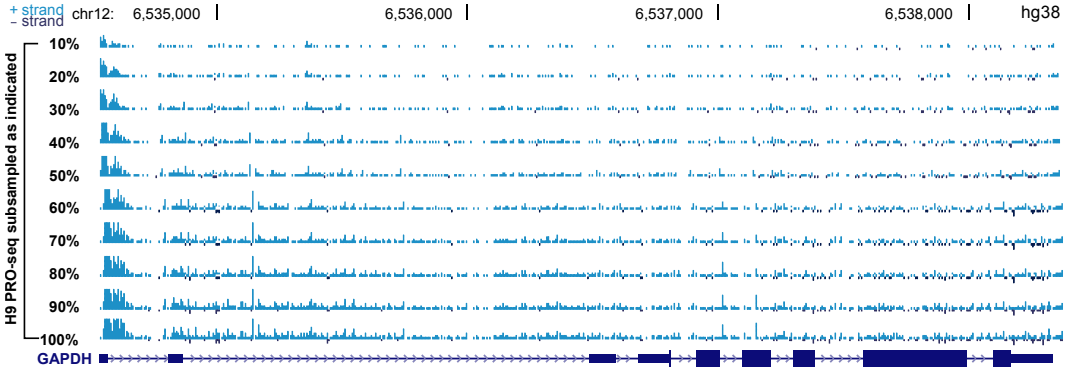

Figure 3.18: **H9 PRO-seq 2 signal tracks show increasing coverage with depth**. Incrementally subsampled H9 PRO-seq 2 library signal tracks display reduced relative coverage at a representative locus (GAPDH). Fixed scale for each track is 10 to -5.

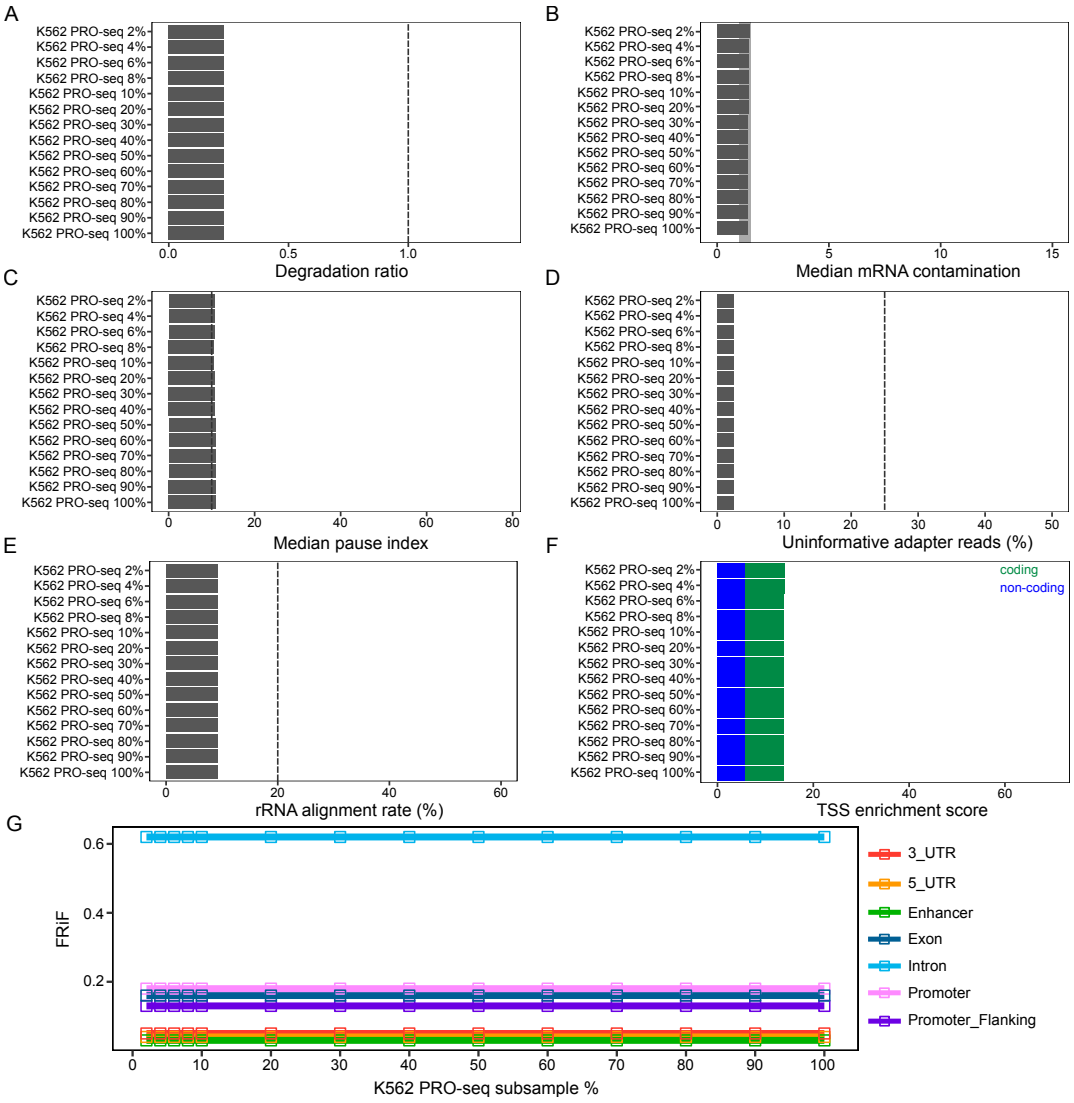

Figure 3.19: **QC metrics are not affected by sequencing depth in subsampled K562 PRO-seq**. Using subsampled K562 PRO-seq data, we show how various metrics behave across a spectrum of sequencing depths: A) Degradation ratio, B) mRNA contamination, C) Pause index, D) the percentage of uninformative adapter reads, E) the rDNA alignment rate, F) and the TSS enrichment scores are unaffected by sequencing depth. G) The FRiF and cumulative FRiF is unaffected by sequencing depth. The complete K562 PRO-seq library (100%) contains approximately 497 million reads.

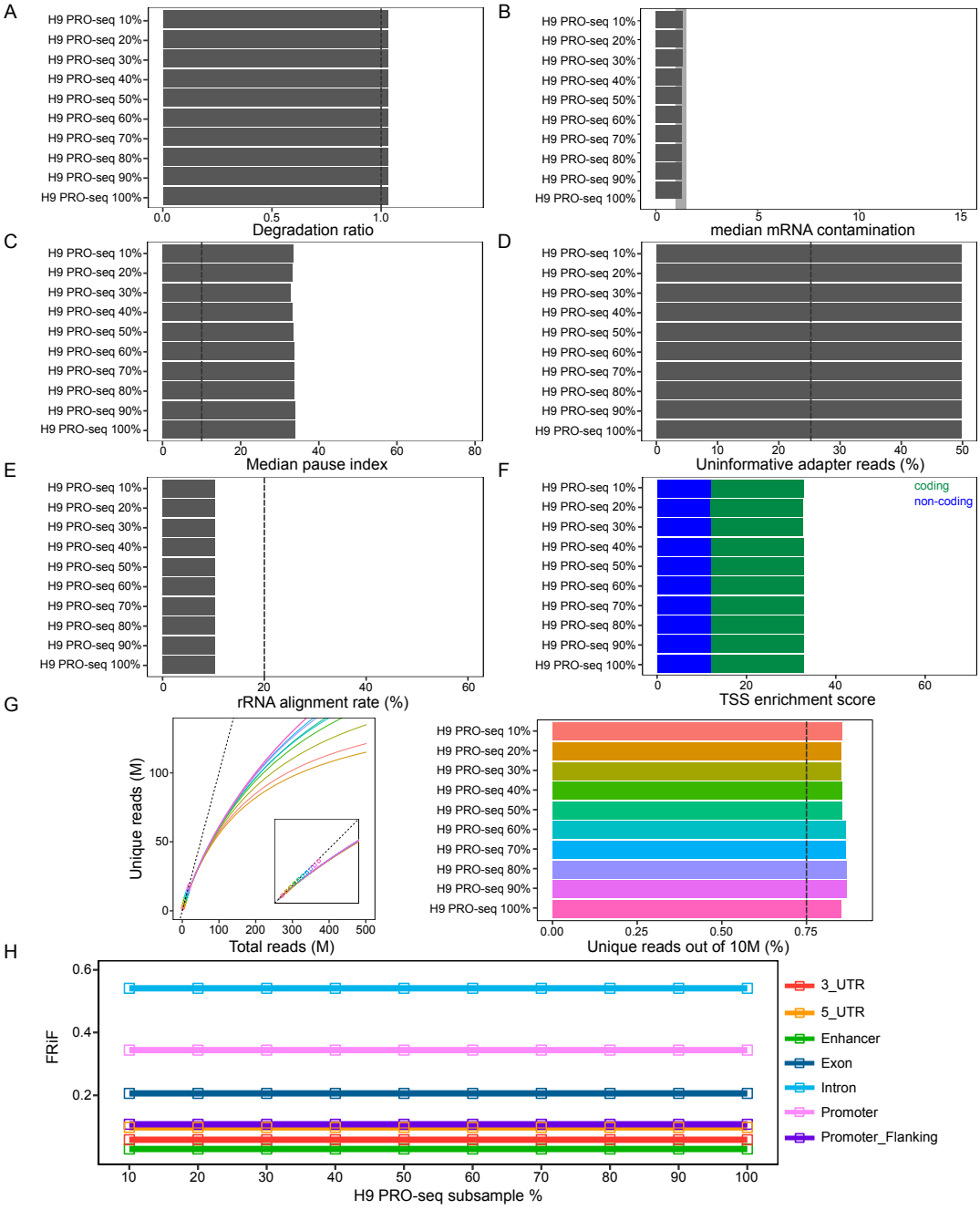

Figure 3.20: **QC metrics are not affected by sequencing depth in subsampled H9 PRO-seq**. Using subsampled H9 PRO-seq data, we show how various metrics behave across a spectrum of sequencing depths: A) Degradation ratio, B) mRNA contamination, C) Pause index, D) the percentage of uninformative adapter reads, E) the rDNA alignment rate, F) and the TSS enrichment scores are unaffected by sequencing depth. G) Library complexity traces plot the read count versus externally calculated deduplicated read counts. Deduplication is a prerequisite, so these plots may only be produced for samples with UMIs. Inset zooms to region from 0 to double the maximum number of unique reads. The position of curves in the left panel at a sequencing depth of 10 million reads (dashed line represents minimum recommended percentage of unique reads). H) The FRiF and cumulative FRiF is unaffected by sequencing depth. The complete H9 PRO-seq library (100%) contains approximately 116 million reads.

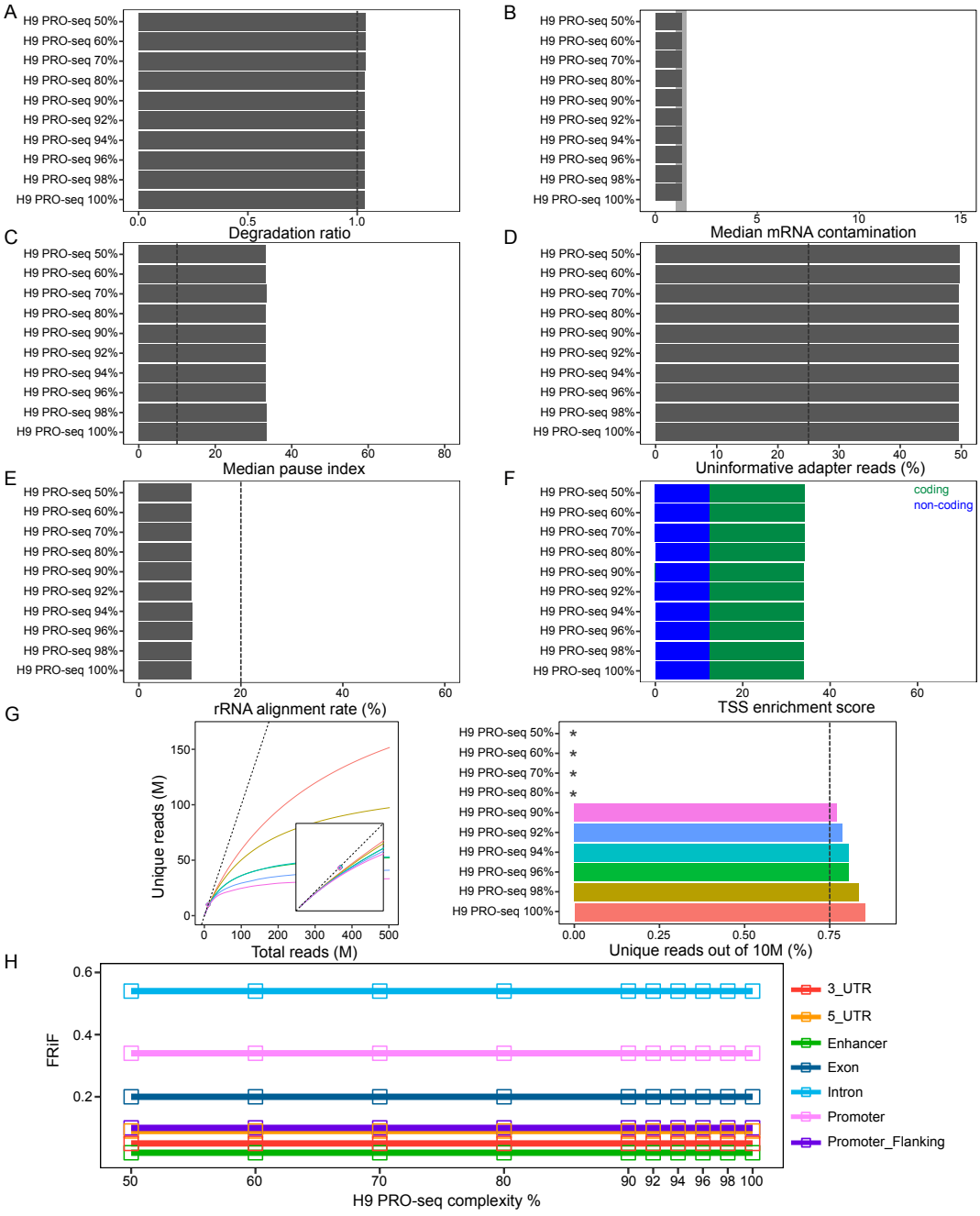

Figure 3.21: **QC metrics are not affected by low library complexity**. Using a synthetic set of libraries, we show how various metrics behave across a spectrum of complexity: A) Degradation ratio, B) mRNA contamination, C) Pause index, D) the percentage of uninformative adapter reads, E) the rDNA alignment rate, F) and the TSS enrichment scores are unaffected by low complexity. G) Library complexity traces plot the read count versus externally calculated deduplicated read counts. Deduplication is a prerequisite, so these plots may only be produced for samples with UMIs. Inset zooms to region from 0 to double the maximum number of unique reads. The right panel represents the position of curves in the left panel at a sequencing depth of 10 million reads (dashed line represents minimum recommended percentage of unique reads). *Libraries with less than 90% uniqueness could not be extrapolated due to saturation. H) The FRiF and cumulative FRiF is unaffected by low complexity. Each library contains 30 million total reads.

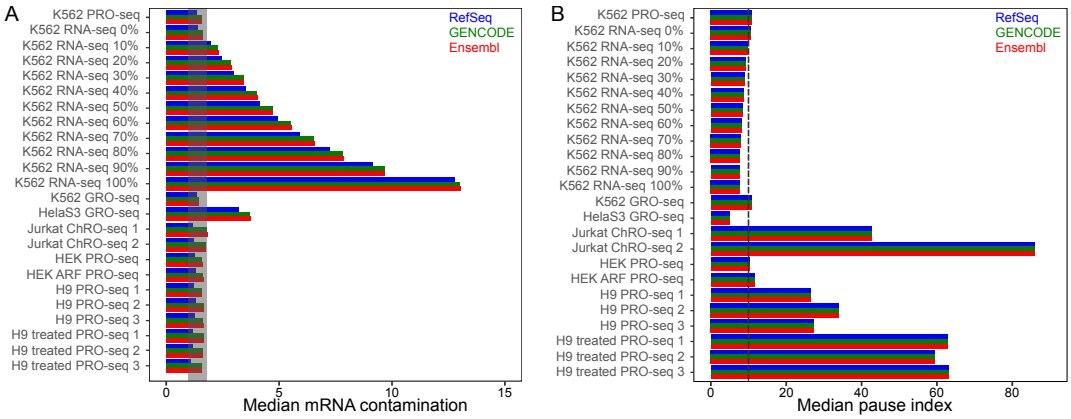

Figure 3.22: **Alternate annotation sources do not affect mRNA contamination and pause index**. A) The mRNA contamination metric and B) the pause index metric are robust across annotations.

4 PEPATAC: An optimized pipeline for ATAC-seq data analysis with serial alignments (modified from [177])

The completion of the nascent-RNA sequencing pipeline provided an improved foundation upon which to base a comprehensive, modular pipeline for ATAC-seq analysis. This ATAC-seq pipeline, which we term `PEPATAC`, is built upon a `Python` infrastructure with a complete, included `R` package for custom consensus peak calling, count table generation, and plotting functions. It enables the analysis of one or many samples with both sample-level and project-level analyses. It was designed to be robust and restartable, highly modular, and to contain extensive documentation to ease adoption. It has since been widely adopted with usage by TCGA [160] and more [224, 330–335].

4.1 Introduction

Because cells package chromatin differently depending on their function and phenotype, profiling chromatin accessibility is a primary experimental approach for understanding cell states [51, 150, 246]. The number of chromatin accessibility experiments has grown dramatically in recent years with the introduction of the assay for transposase-accessible chromatin (ATAC-seq) [170]. With ATAC-seq now widespread, there is demand for analytical approaches [213, 214], including systematic processing pipelines to facilitate the goal of reproducible research and ease cross-study comparisons [336, 337].

To address this need we developed `PEPATAC`, a fast and effective ATAC-seq pipeline that easily generalizes across compute contexts and research environments. This pipeline has been built over years of experience analyzing chromatin accessibility experiments and implements several concepts that make it effective. These include ATAC-specific quality control outputs, both nucleotide-resolution and smoothed signal tracks, and a serial alignment strategy to deal with high mitochondrial contamination. Our serial alignment strategy, or 'prealignments,' allows the user to configure a series of genomes to align to before the primary genome. `PEPATAC` provides a framework that allows a user to align serially in customized order to as many genomes as desired, which will be useful for many situations, including species contamination, dual-species experiments, repeat model alignments, decoy contamination, or spike-in controls.

While numerous ATAC-seq pipelines exist [213, For more in-depth coverage see: 214], `PEPATAC` is designed with modularity and flexibility as paramount design considerations (Fig. 4.1a). `PEPATAC` is compatible with the Portable Encapsulated Projects (PEP) format [303], which defines a common project metadata description, allowing projects that use `PEPATAC` to be

easily analyzed using any PEP-compatible tool. It also provides the possibility for a single project description to be shared across pipelines, computing environments, and analytical teams. `PEPATAC` is easily customizable, including changing individual command settings or even swapping specific software components by modifying a few lines of human readable configuration files.

`PEPATAC` does not rely on any specific local or cloud computing infrastructure, and it has already been deployed successfully in various compute environments at multiple research institutes to yield numerous peer-reviewed studies [160, 332, 338–340]. While *all* ATAC-seq pipelines use several common bioinformatic tools (Fig. 4.4), we simplify the creation of a computing environment with the required command-line tools using conda [341], or either docker or singularity with the bulker multi-container environment manager [304].

`PEPATAC` includes a well-documented code base with detailed installation instructions, tutorials, and example projects, so it is useful for both the bench biologist and bioinformatician alike. We anticipate that this pipeline will provide a useful complete analysis for basic ATAC-seq projects and serve as a unified starting point for more advanced ATAC-seq projects.

4.2 Materials and Methods

4.2.1 PEPATAC configuration

a

	flexible trimmer	prealignments	flexible aligner	smooth tracks	base-resolution tracks	flexible peak caller	HTML report	peak annotation	read annotation	motif analysis	containerized	flexible computing	differential analysis
AIAP	×	×	×	✓	×	×	✓	✓	✓	×	✓	✓	✓
ATAC2GRN	×	×	×	×	×	×	×	×	×	×	✓	✓	×
ATAC-pipe	×	×	×	✓	×	×	×	×	×	×	×	✓	✓
ATACProc	×	×	×	✓	×	×	✓	×	×	✓	×	×	×
CIPHER	×	×	✓	✓	×	✓	✓	×	×	×	✓	✓	×
ENCODE	✓	×	✓	✓	×	×	✓	✓	✓	×	✓	✓	×
esATAC	×	×	×	✓	×	×	✓	✓	×	✓	×	×	×
GUAVA	×	×	×	✓	×	×	×	×	✓	×	×	✓	✓
I-ATAC	×	×	×	×	×	×	×	×	×	×	×	✓	×
nfcore/atacseq	×	×	×	✓	×	×	✓	✓	✓	×	✓	✓	✓
PEPATAC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×
pyflow-ATAC-seq	×	×	×	✓	×	×	✓	×	×	×	×	✓	×
seq2science	✓	×	✓	✓	×	✓	✓	×	×	×	×	✓	✓
snakePipes	×	×	×	✓	×	✓	✓	×	×	×	×	✓	✓
Tobias Rausch	×	×	×	✓	×	×	✓	✓	✓	✓	✓	×	×

b

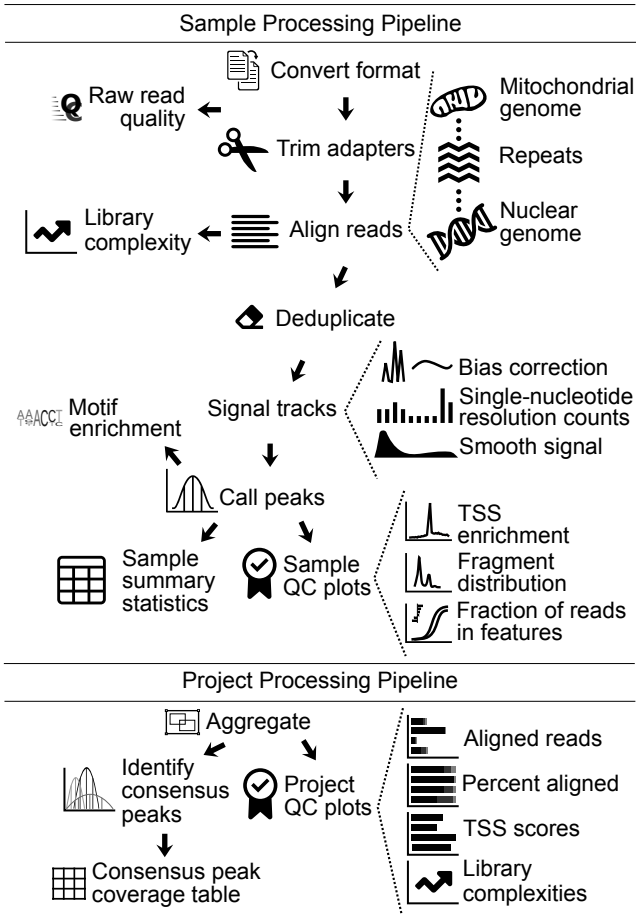

Figure 4.1: **`PEPATAC` is feature-rich with a logical workflow**. (a) We compared features across 14 ATAC-seq pipelines (AIAP [342]; ATAC2GRN [343]; ATAC-pipe [344]; ATACProc [345]; CIPHER [346]; ENCODE [347]; esATAC [348]; GUAVA [349]; I-ATAC [350]; nfcore/atacseq [351]; pyflow-ATAC-seq [352]; seq2science [353]; snakePipes [354]; Tobias Rausch [355]) and `PEPATAC` stands out for being feature-rich . (b) Reads are preprocessed, serially aligned to the mitochondrial genome, curated repeats, and then the nuclear genome. `PEPATAC` generates both smooth and exact signal plots, called peaks, and QC output plots and tables.

The PEPATAC pipeline is divided into two major parts (Fig. 4.1b): First, it processes each sample individually at the *sample level*. Once sample processing is complete, the *project-level* part aggregates, analyzes, and summarizes the results across samples. PEPATAC is composed of two primary Python scripts that may be run from the command-line. Sample information and parameters are passed to the pipeline as command-line arguments (see `pepatac.py --help`), making it simple to use as a standalone pipeline for individual samples without requiring a complete project configuration. Project level output is produced using the project level pipeline (see `pepatac_collator.py --help`). PEPATAC is built using the Python module `pypiper` [356], which provides restartability, file integrity protection, copious logging, resource monitoring, and other features. Individual pipeline settings can also be configured using a pipeline configuration file (`pepatac.yaml`), which enables a user to specify absolute or relative paths to installed software, change adapter input files for trimming, and parameterize alignment and peak calling software tools. This configuration file comes with sensible defaults and will work out-of-the-box for research environments that include required software in the shell PATH, but it also may be configured to fit any computing environment and adapt to project-specific parameterization needs.

4.2.2 Refgenie reference assembly resources

Like any genome analysis, PEPATAC relies on reference genome annotations. To ensure that results are comparable across runs, it's important to use the same reference assembly. To manage these assets in a reproducible and robust manner, PEPATAC uses *refgenie*. Refgenie is a reference genome assembly asset manager that simplifies access to pre-indexed genomes and annotations for common assemblies, and also allows generating new standard reference genomes or annotations as needed while maintaining asset provenance [322, 357]. For a complete analysis, PEPATAC requires several `refgenie`-managed assets: fasta, chrom_sizes, bowtie2_index, blacklist, refgene_tss, and feat_annotation. These can be either downloaded automatically or built manually, which require a genome fasta file, a gene set annotation file from RefGene, and an Ensembl gene and regulatory build annotation file. Using PEPATAC with `seqOutBias` requires the additional `refgenie` tallymer_index asset built for the same read length as the data. Many of these assets may also be directly specified at the command line should a user not have `refgenie`-managed versions available. The TSS annotation file, region blacklist, and feature annotation file may all be specified to use a local, user-specified file. For example, while ENCODE provides a common set of regions that are aberrantly overrepresented in sequencing experiments (e.g. a blacklisted set of regions) [358], a user may create their own version of regions that should be excluded from consideration and point to

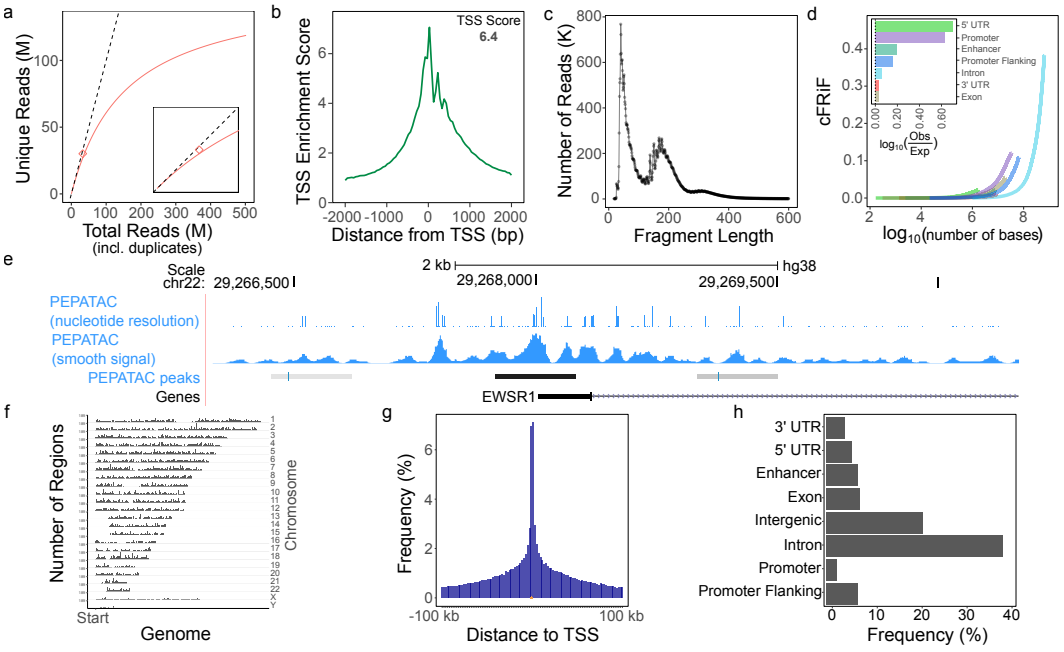

Figure 4.2: **Example `PEPATAC` QC plots for reads and peaks**. (a) Library complexity plots the read count versus externally calculated deduplicated read counts. Red line is library complexity curve for SRR5427743. Dashed line represents a completely unique library. Red diamond is the externally calculated duplicate read count. (b) TSS enrichment quality control plot. (c) Fragment length distribution showing characteristic peaks at mono-, di-, and tri-nucleosomes. (d) Cumulative fraction of reads in annotated genomic features (cFRiF). Inset: Fraction of reads in those features (FRiF). e) Signal tracks including: nucleotide-resolution and smoothed signal tracks. `PEPATAC` default peaks are called using the default pipeline settings for MACS2 [255]. (f) Distribution of peaks over the genome. (g) Distribution of peaks relative to TSS. (h) Distribution of peaks in annotated genomic partitions. Data from SRR5427743.

this file manually.

4.2.3 File inputs and adapter trimming

`PEPATAC` sequentially trims, aligns, and analyzes sequences (Fig. 4.1b). `PEPATAC` accepts sequence data input in 3 formats: unaligned BAM, separated FASTQ, or interleaved FASTQ format. The pipeline first converts the input format into FASTQ (if necessary) for adapter trimming. For adapter trimming, users may select between `skewer` [359], `trimmomatic` [360], or an included `Python` tool using command-line arguments or the PEP configuration file. The pipeline stores quality control results including the number of raw, trimmed, or duplicated reads, and runs `FastQC` [361] if installed.

4.2.4 Prealignments and mitochondrial DNA

Because ATAC-seq data can have a high proportion of reads mapping to the mitochondrial genome (from 15%-50% in a typical experiment up to 95% in some experiments [362]), we considered how to optimize the pipeline to deal with abundant mitochondrial DNA (mtDNA). High mtDNA exacerbates the alignment challenge caused by nuclear-mitochondrial DNA (NuMts), which are mtDNA sequences that have integrated into the nuclear genome throughout eukaryotic evolution [364]. NuMts represent nonfunctional, truncated, and mutation-ridden copies of mitochondrial protein-coding genes; therefore, we assume that ATAC reads mapping to them are highly likely to be erroneous alignments. The typical strategy is to align to the mitochondrial and nuclear genomes simultaneously, and then remove nuclear-mitochondrial DNA (NuMts) post-hoc using a blacklist, but this suffers from three disadvantages: First, it is inefficient to align lots of mtDNA to the larger nuclear genome; second, reads that match both NuMt and mtDNA will be (incorrectly) split between the two, and third, this approach relies on an accurate pre-constructed annotation of NuMt locations, which may not be available for every reference genome. Furthermore, due to mitochondrial genetic diversity within and across cells, some reads derived from true mtDNA may in fact map better to the reference NuMt than to the reference mtDNA sequence. Also, reads that span the artificial breakpoint in the linear mtDNA reference may find an adequate NuMt match, but would never align to the mtDNA.

We found that by separately aligning first to the mitochondrial genome, we alleviated the challenges with simultaneous alignments. To capture NuMts that span the artificial breakpoint induced by converting the circular mitochondrial DNA into a linear representation for alignment, we use a doubled mitochondrial reference sequence, which enables non-circular aligners to align reads that span the breakpoint. By default, the pipeline is configured to

align reads first to the doubled mitochondrial reference genome, but may be easily configured to perform any number of additional serial alignments.

4.2.5 Alignments, deduplication, and library complexity

For prealignments and primary alignment, `PEPATAC` employs `bowtie2` by default [325]. `Bowtie2` settings are configurable in the pipeline configuration file but come with sensible defaults of `-k 1 -D 20 -R 3 -N 1 -L 20 -i S,1,0.50` for prealignments and `--very-sensitive -X 2000` for nuclear genome alignment. Users may optionally use `bwa` [365] with settings similarly configurable in the pipeline configuration file (default: `-M`). Following alignment, reads with mapping quality scores below 10 and any residual mitochondrial reads are removed and read deduplication is carried out using `samblaster` [366], but picard's `MarkDuplicates` [367], or `samtools` [328] may also be utilized based on user preference. `PEPATAC` utilizes *preseq* [312] to calculate and plot sample library complexity at the current depth, and includes the number of independently calculated duplicates (Fig 4.2a). The pipeline also projects the unique fraction of the library at 10M total reads. These metrics provide an estimate of library complexity and allow the user to determine the value of subsequent sequencing.

4.2.6 Library QC metrics

For quality control, `PEPATAC` provides a TSS enrichment plot, produced by aggregating reads present in regions 2000 bases upstream and downstream of a reference set of TSSs (Fig 4.2b). Enrichment is calculated as the average number of reads in a 100 bp window around the TSS divided by the average number of reads in the first 200 bases of the entire region. This yields low signals in the tails with a peak in the center, which we take to be the TSS enrichment score. `PEPATAC` also produces a fragment length distribution plot (Figure 4.2c). A standard quality ATAC-seq library is expected to yield clearly defined peaks at open chromatin (<100bp), mononucleosomes (200 bp), and sequentially smaller peaks representing multi-nucleosomes at regular intervals. To evaluate the enrichment of all reads across genomic partitions, `PEPATAC` plots both the fraction and cumulative fraction of reads (FRiF, cFRiF respectively) in genomic features (Fig 4.2d). A novel feature of `PEPATAC` includes the plotting of the fraction of reads in any feature type, not solely in peaks. This is plotted as the cumulative sum of reads in each feature divided by the total number of aligned reads against the cumulative sum of bases in each feature. The relative proportion of each feature can be then be directly compared. The standard feature annotation produced and managed by `refgenie` includes Ensembl defined enhancers, promoters, promoter flanking regions, 5' UTR, 3' UTR, exons, and introns in

that order. Users can specify an alternative annotation file, either a custom one or simply a different sort order, using the `--anno-name` pipeline parameter. For a quality sample, the proportion of reads in peaks should be the most enriched, reflecting the specificity of the peak calls for that sample.

4.2.7 Signal tracks and peak calling

Alignments are used to generate two signal tracks: one that records the exact location of transposition events, and one that is smoothed (Fig 4.2e). These tracks may be used for different downstream analyses; the exact track is useful for analysis that requires nucleotide-resolution, while the smoothed version is often preferred for visualization and peak analysis. Reads, representing transpoase cut-sites, are extracted from the deduplicated, low-quality removed, primary genome mapped BAM file into a wiggle-like track. For the exact signal track, these cut-sites are shifted +4 bases for positive strand reads and -5 bases for negative strand reads. For the smooth signal track, we extend the shifted exact sites +/- 25 bases to yield 50 bp smoothed windows around the exact cut-site position. `seqOutBias` is an optional tool that can be used to correct for enzymatic (e.g. Tn5 transposase) bias and generate tracks for visualization [311]. The bias itself is corrected using a k-mer mask for the plus and minus strand Tn5 recognition sites and by taking the ratio of genome-wide observed read counts to the expected sequence based counts for each k-mer [311]. The k-mer counts take into account mappability at a given read length using GenomeTools' `Tallymer` program [329].

An earlier study found multiple peak callers worked well with chromatin accessibility data [368], and `PEPATAC` provides the option to use `F-Seq` [254], `MACS2` [255], `Genrich` [257], `HOMER` [267], or `HMMRATAC` [258] for peak calling, with parameters customizable in the pipeline configuration file. `MACS2` is used by default (`--shift -75 --extsize 150 --nomodel --call-summits --nolambda --keep-dup all -p 0.01`). The default settings are intended to maximize recall and sensitivity. More stringent settings can be easily adopted by modifying the pipeline configuration file. Called peaks are standardized by extending up and down 250 bases (a tunable parameter, `--extend`) from the summit of each peak to establish peaks 500 bases in width. Any peaks which then extend beyond chromosome boundaries are trimmed. Utilizing fixed-width peaks reduces bias towards larger peaks in both count-based and motif analyses while simultaneously improving the identification of consensus peak sets by reducing the likelihood of extraordinary large peaks created through the union and merging of multiple peak sets. Finally, peak scores are normalized to score per million by dividing by the sum of scores over 1M.

PEPATAC also produces several plots detailing enrichment of reads in peaks including: the distribution of peaks across the genome by chromosomal location (Fig 4.2f), the distribution of peaks relative to TSSs (Fig 4.2g), and the distribution of peaks within genomic partitions (Fig 4.2h). The TSS distance distribution shows the distance of called peaks with respect to TSSs grouped in log-scale bins. Finally, users may optionally employ HOMER to calculate motif enrichments in called peaks [369].

4.2.8 Running multiple samples with PEPATAC

To run the pipeline across multiple samples in a larger project, the pipeline uses the job submission engine looper [370], which employs the Portable Encapsulated Project standardized definition of project metadata [303](Fig. 4.5). This standard project format enables a pipeline to be run on any project that follows the format, which is simple, standardized, and well-documented. Looper enables the PEPATAC pipeline to be run in any compute environment, including locally (the default) on a single laptop or desktop, or with any cluster resource manager. It also can be used with containers. Additionally, looper's project format gives pipeline users access to APIs written in Python and R for downstream analysis of pipeline results.

For the user whose environment is set up to run containers, we enable container use with either Docker or Singularity via a single image file or through the multi-container environment manager, bulker [304]. Using bulker, PEPATAC may be run in containers across samples and compute environments, simplifying deployment by requiring only bulker and the PEPATAC pipeline itself, eliminating the need to install each required package independently.

4.2.9 Aggregating results from multiple samples

To summarize and incorporate data across samples, the second step in a PEPATAC analysis is to run a project-level pipeline (pepatac_collator.py) that identifies consensus peaks across a project and calculates sample coverage of those consensus peaks in a convenient table for easy downstream analysis. To establish consensus peaks, PEPATAC identifies overlapping (1 bp, a tunable parameter: --min-olap) peaks between every sample in a project and defines the consensus peak's coordinates based on the overlapping peak with the highest score. Peaks present in at least 2 (parameter: --cutoff) samples with a minimum score per million greater than or equal to 5 (parameter: --min-score) are retained. A peak count table is then provided where every sample peak set is overlapped against the consensus peak set. Individual peak counts for an overlapping peak are weighted by multiplying by the percent overlap of the sample peak with the consensus peak.

For navigating results, PEPATAC provides both sample and project level reports in a convenient, easy-to-navigate HTML report with project-level summary table and plots, job status page, and individual sample pages with sample statistics and QC plots all at your fingertips. In addition, looper will produce summary plots from individual sample statistics including the number of aligned reads, percent aligned reads, TSS scores, and library complexities. A user can produce the HTML report during a run or after completion, with the job status page providing information on whether a sample has failed, is still running, or has already completed.

4.3 Results

To demonstrate PEPATAC's default workflow and output, we analyzed samples from the original standard ATAC [170], fast ATAC [158], and omni ATAC [174] protocol papers. This dataset includes human ATAC-seq reads from 33 standard ATAC, 152 fast ATAC, and 139 omni ATAC samples (Supplemental file 1). PEPATAC provides output and quality control results both for individual samples and for the project as a whole. For each sample, PEPATAC produces narrowPeak and bigWig files to visualize nucleotide-resolution alignments, smoothed alignments, and peak calls. PEPATAC also produces summary statistics files that report the number of reads, duplicates, genome alignment rates, transcription start site (TSS) enrichment score, number of called peaks, fraction of reads in peaks (FRiP), and job runtime among others for every sample in a project.

4.3.1 Performance

PEPATAC is designed to be computationally efficient. To evaluate how PEPATAC scales with increasing numbers of reads, we ran 430 ATAC-seq samples of varying input size through PEPATAC (Supplemental file 4). We then placed samples in 500MB input file size bins and compared runtimes and peak memory usage (Fig. 4.6). Runtime scales linearly with increasing file size, but importantly, even samples with more than 150 million reads completed in less than 8 hours (Fig 4.6a). We also show that PEPATAC, with default settings, only utilizes between 5-9 GB at peak memory use (Fig 4.6b).

4.3.2 Prealignments

To evaluate the advantage of serially aligning to the mitochondrial genome (Fig. 4.3a), we measured the total alignment runtime of synthetic mixtures of mitochondrial-aligning (mtDNA) and whole human-aligning (hg38) sequences with and without prealignments. We constructed libraries of mixed mtDNA:hg38 mapping ATAC-seq reads from 0% to 100%

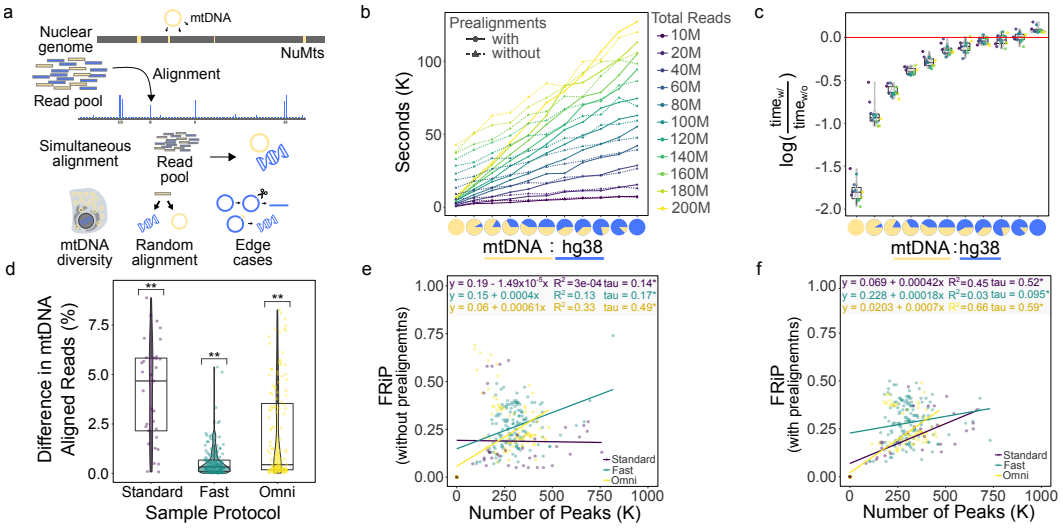

Figure 4.3: **PEPATAC prealignments increase mapped mtDNA reads, improve computational efficiency, and positively influences the fraction of reads in peaks (FRiP) metric.** (a) NuMTs represent a significant complication of simultaneous alignment. (b) At mtDNA percentages from 10-100% at total read numbers ranging from 10-200M, using prealignments dramatically reduces run time. (c) Log ratio of prealignments runtimes versus no prealignment runtimes yields significant savings. (d) There is a significant increase in the percent of reads mapped to mitochondrial sequence when using prealignments versus not across standard, fast, and omni-ATAC protocols. (e) As reported for ChIP-seq [371], FRiP is positively correlated with the number of called peaks. (f) With prealignments, the positive correlation between FRiP and the number of called peaks tends to increase ((d) ** = $p < 0.001$; t-test (mu = 0) with Benjamini-Hochberg correction. (e-f):* = $p < 0.0001$; Kendall rank correlation coefficient).

mtDNA in increments of 10%, at 10 million, 20 million, and up to 200 million total reads in increments of 20 million reads, resulting in 121 different library combinations. We recorded the alignment time for each input file with and without prealignments (Fig. 4.3b). To determine for which scenarios using prealignments is beneficial, we calculated the log ratio of run times with prealignments versus without prealignments and found that using prealignments reduces the total time of alignment even when mtDNA alignment rates are under 10% (Fig. 4.3c). In addition to speed and efficiency gains, `PEPATAC` with prealignment compared to without prealignment to mtDNA yields higher alignment rates to mitochondrial sequence than aligning to a combined human and mitochondrial genome as is commonly performed (Fig. 4.3d). This is true for every sample tested no matter the library preparation protocol nor percent mitochondrial contamination (Fig. 4.7). This result indicates that the common approach of simultaneously aligning to the nuclear and mitochondrial genomes systematically underestimates the fraction of mitochondrial reads in an experiment. We therefore propose that mitochondrial alignment rates are generally underestimated by about 1-5% in published reports.

To show how prealignments successfully depletes reads aligning to NuMTs, we ran a standard ATAC (SRR5427804), fast ATAC (SRR2920492), and omni ATAC (SRR5427806) sample through `PEPATAC` with no prealignments, prealignment to mitochondrial sequence, and prealignment to mitochondrial, ribosomal, and known repeat sequences. We then compared the highest signal peaks between each prealignment strategy across each ATAC-seq protocol. We used `BLAST` [372] to annotate the highest signal peaks and then intersected called peaks under each strategy with the ENCODE blacklist [358], which normally is used to filter results in `PEPATAC` by default. The omni ATAC sample had the least number of aberrant high signal peaks with only a single NuMT peak identified in the top 10 highest signal peaks and *only* present when analyzed without prealignments. Significantly, as soon as mitochondrial prealignment is included, this peak is excluded (Supplemental file 3, Fig. 4.8a). Of the top 100 omni ATAC peaks, there are fewer overlaps with blacklisted regions, both overall, and as we increase the number of prealignments. With no prealignments there are 4 blacklisted regions in the top 100 and only 2 with prealignments (Supplemental file 3). As omni ATAC is reported to reduce mitochondrial reads, this result is expected. Furthermore, this difference is highlighted as we compare both fast ATAC and standard ATAC. Three of the top 10 peaks from the fast ATAC sample without prealignments aligned to mitochondrial sequence (Supplemental file 3). These are eliminated with prealignments. Additionally, without prealignments, 22 of the top 100 peaks intersect blacklisted regions. Only 18 overlap with mitochondrial prealignment, and significantly, only 3 of the top 100 overlap blacklisted regions when prealigning includes

ribosomal and repeat regions (i.e. satellite DNA). This suggests that a number of regularly identified peaks should typically be excluded in the absence of prealignments. While a blacklist does an excellent job at removing these regions, prealignment achieves similar results while also removing additional non-blacklisted regions that are likely spurious (mapping to unmapped regions or to different species, see Supplemental file 3). These results are even more obvious with standard ATAC. Standard ATAC without prealignment to mitochondria mapped 8 of the top 10 peaks to NuMTs (Supplemental file 3). These are removed with prealignment to mitochondria. Furthermore, the number of blacklisted regions drops from 17 without prealignments to 7 with mitochondrial prealignment and only 2 with mitochondrial, ribosomal, and repeat region prealignment. Because prealignment reduces spurious peak assignment (Supplemental file 3, Fig. 4.8b) *and* it reduces total runtime in nearly every scenario (Fig. 4.3c), prealignment is an effective strategy to include in every pipeline run.

4.3.3 Peak caller comparison

To evaluate the difference in called peaks when using different peak callers, we compared both the `PEPATAC` determined consensus peaks and the peaks from a single sample (SRR5210416) produced when using different peak callers (`Fseq`, `Genrich`, `HOMER`, `HMMRATAC`, `MACS2` with variable peaks, and `MACS2` with fixed peaks). Similarity between the intervals was evaluated with a modified Jaccard statistic [373] implemented in the `bedtools` [323] package. At the single sample level `MACS2` with variable peak width is the most similar in output to `MACS2` with fixed peaks and `Fseq` (Fig. 4.9a, see Supplemental file 2). Interestingly, the least similar peak results are from `Genrich` and `HMMRATAC`, which possibly reflects the goal of both tools being designed to evaluate ATAC-seq data as opposed to originally being developed for ChIP-seq (Fig. 4.9a). These differences become more pronounced at the consensus peak level, with `HMMRATAC` becoming more dissimilar (average jaccard statistic = 0.31, Supplemental file 2) to the other peak callers (Fig. 4.9b).

We also asked whether this difference was due to an improvement in reduced peak calling at nuclear mitochondrial sequences (NuMTs), repeat regions, or high signal regions. One way to evaluate this is to determine the number of intersections of the individual peak caller called regions against a known blacklist [358] and to BLAST [372] the highest signal peaks. Indeed, HMMRATAC overlaps the least number of blacklisted regions (231 versus the maximum of 756 with HOMER; see Supplemental file 2) and it turns out a number of both the blacklisted regions and the highest signal peaks are NuMTs or repeat regions (Supplemental file 3). While MACS2 remains the most commonly employed peak caller across ATAC-seq pipelines, further comparative studies may better illustrate the utility of some of the more recently developed

peak callers.

4.3.4 Library QC comparison

Several of the QC metrics (e.g. TSS enrichment score, the fragment distributions, non-redundant fractions, and the PCR bottlenecking coefficients 1 and 2) employed by `PEPATAC` are near-universal in the field, and as such are calculated in the same manner. To evaluate how different annotations may affect the TSS score, we also compared TSS annotations from Ensembl, Gencode, and Refgene (`PEPATAC` default). Refgene produces higher TSS scores (Fig. 4.10), which reflects the fact that Refgene contains only the most commonly employed transcription start sites for each gene whereas both Ensembl and Gencode include all known sites, diluting the aggregated signal.

4.3.5 Fraction of reads in peaks

It has also been reported that in ChIP-seq experiments, but not specifically in ATAC-seq, that FRiP correlates positively with the number of identified peaks [371] (Fig. 4.3e). In libraries with significant mitochondrial contamination, for example, from libraries produced using standard-ATAC library preparation protocols, this correlation is emphasized when using prealignments (Fig. 4.3f). We next sought to understand how the serial alignment strategy affects calculation of Fraction of Reads in Peaks (FRiP). FRiP is a common qualitative measure of enrichment and sample quality. However, FRiP calculations are poorly defined, making it dangerous to compare FRiP scores among different protocols and approaches. ENCODE defines the denominator of the FRiP score to be total mapped reads (ENCODE Terms). If only one genome is used for alignment, then the calculation is clear, but for a serial alignment pipeline, the FRiP score depends on whether the denominator includes reads mapped to the nuclear genome only, or to all genomes (Fig. 4.11c,d). By default, `PEPATAC` uses the deduplicated, low-quality removed, primary genome mapped BAM file to calculate the fraction of reads in the final called peak output file, which by default utilizes fixed width peaks and has removed any blacklisted regions. This has the consequence of changing the FRiP calculation based on whether prealignments were used (Fig. 4.11c,d). When using prealignments, the default FRiP calculation will significantly increase, because the number of reads mapped to the primary genome is reduced due to reads mapping more accurately to the mitochondrial genome and thus being excluded from downstream analysis. When FRiP is calculated using the total mapped reads (prealignments **and** primary alignment), these relationships are inversed (Fig. 4.11c,d). In any scenario, prealignments lead to more total mapped reads, due to more efficient mitochondrial alignment. As more recent ATAC-seq sample preparation protocols

intentionally reduce mitochondrial contamination, these differences are most pronounced when using the original, standard ATAC-seq protocol. Therefore, reliance on a specific cutoff (e.g. 0.3 or greater) as indicative of a quality sample must be relative to protocol and method.

4.4 Discussion

`PEPATAC` is an efficient, user-friendly ATAC-seq pipeline that produces helpful quality control plots and signal tracks that provide a comprehensive starting point for further downstream analysis. Two key benefits of the `PEPATAC` pipeline over existing pipelines are its flexibility and modularity. `PEPATAC` is uniquely flexible, for example, by allowing pipeline users to serially align to multiple genomes, to select from multiple aligners, peak callers, and adapter trimmers, while providing a convenient, configurable interface so a user can adjust parameters for individual pipeline tasks. Furthermore, `PEPATAC` reads projects in PEP format, a standardized, well-described project definition format, providing a reproducible interface with Python and R APIs to simplify downstream analysis.

Because `PEPATAC` is built on `looper`, it is easily deployable on any compute infrastructure, including a laptop, a compute cluster, or the cloud. It is thereby inherently expandable from single to multi-sample analyses with both project level and individual sample level quality control reporting. This means that a user may submit any number of samples using *a single looper command and corresponding PEP metadata file.* Its design allows for simple restarts at any step in the process should the pipeline be interrupted. Due to its modular construction multiple software options for primary pipeline steps are available, creating a swappable pipeline flow path with individual steps adaptable to future changes in the field. `PEPATAC` is a rapid, flexible, and portable ATAC-seq project analysis pipeline providing a standardized foundation for more advanced inquiries.

4.4.1 Documentation and links

- `PEPATAC` v0.9.16: pepatac.databio.org.
- PEP metadata standards: pep.databio.org.
- `Looper` job submission engine: looper.databio.org.
- `Refgenie` reference genomes: refgenie.databio.org.
- Source code to reproduce output for this paper: github.com/databio/pepatac_paper_data.

4.5 Supplemental

4.5.1 Supplemental figures

	trimming	read QC	mapping	deduplication	filtering	signal generation	peak calling	QC	downstream analyses
AIAP	cutadapt	FastQC	bwa	picard	samtools methylQA	UCSC tools	MACS2	MultiQC	DESeq2
ATAC2GRN	NA	NA	bowtie2	NA	NA	NA	HOMER	NA	HINT
ATAC-pipe	custom python	custom python	bowtie2	picard	samtools	UCSC tools	MACS2	custom python	CENTIPEDE DESeq2 HOMER
ATACProc	trim_adapters.py*	NA	bowtie2	picard DeepTools	samtools DeepTools	UCSC tools DeepTools	MACS2	ataqv	HINT-ATAC HOMER DeepTools custom python
CIPHER	BBDUK	FastQC	bbmap bowtie2 bwa hisat2 star	NA	samtools	DeepTools	MACS2 epic	MultiQC	NA
ENCODE	trimmomatic cutadapt	NA	bowtie2 bwa	picard	samtools bedtools	UCSC tools	MACS2	custom code	IDR
esATAC	AdapterRemoval	NA	Rbowtie2	custom R	NA	custom R	F-Seq	custom R	ChIPpeakAnno
GUAVA	cutadapt	FastQC	bowtie2	NA	NA	UCSC tools	MACS2	custom code	DESeq2 ChIPpeakAnno
I-ATAC	trimmomatic	FastQC	bwa	picard	NA	NA	MACS2	NA	NA
nfcore/atacseq	Trim Galore!†	FastQC	bwa	picard	samtools bedtools pysam bamtools	bedtools UCSC tools	MACS2	ataqv	DESeq2
PEPATAC	skewer trimmomatic trim_adapters.py‡	FastQC	bowtie2 bwa	samblaster picard samtools	samtools bedtools	custom python	MACS2 F-Seq2 Genrich HMMRATAC HOMER	custom code	HOMER custom code
pyflow-ATAC-seq	atactk§	FastQC	bowtie2	samblaster	samtools	DeepTools	MACS2	ataqv MultiQC	CENTIPEDE
seq2science	Trim Galore!† fastp	FastQC	bowtie2 bwa hisat2 star	picard	samtools	DeepTools	MACS2 Genrich HMMRATAC	MultiQC	custom code
snakePipes ATAC-seq	cutadapt	FastQC	bowtie2	sambamba	samtools	DeepTools	MACS2 Genrich HMMRATAC	MultiQC	CSAW
Tobias Rausch	cutadapt	FastQC	bowtie2	biobambam2	samtools	Alfred	MACS2	Alfred	HOMER custom R tutorial
OVERALL	cutadapt	FastQC	bowtie2	picard	samtools	UCSC tools	MACS2	MultiQC	HOMER DESeq2

Figure 4.4: **ATAC-seq pipelines universally require several common bioinformatic tools.** While all pipelines require a number of common bioinformatic tools, PEPATAC offers the greatest flexibility and includes a number of the most popular tools.

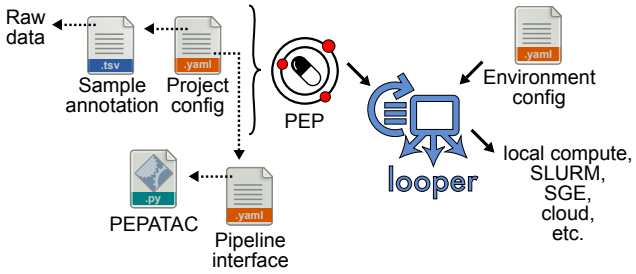

Figure 4.5: **Deploying PEPATAC across multiple samples using looper**. The PEPATAC pipeline can be easily run across multiple samples in any computing environment using looper.

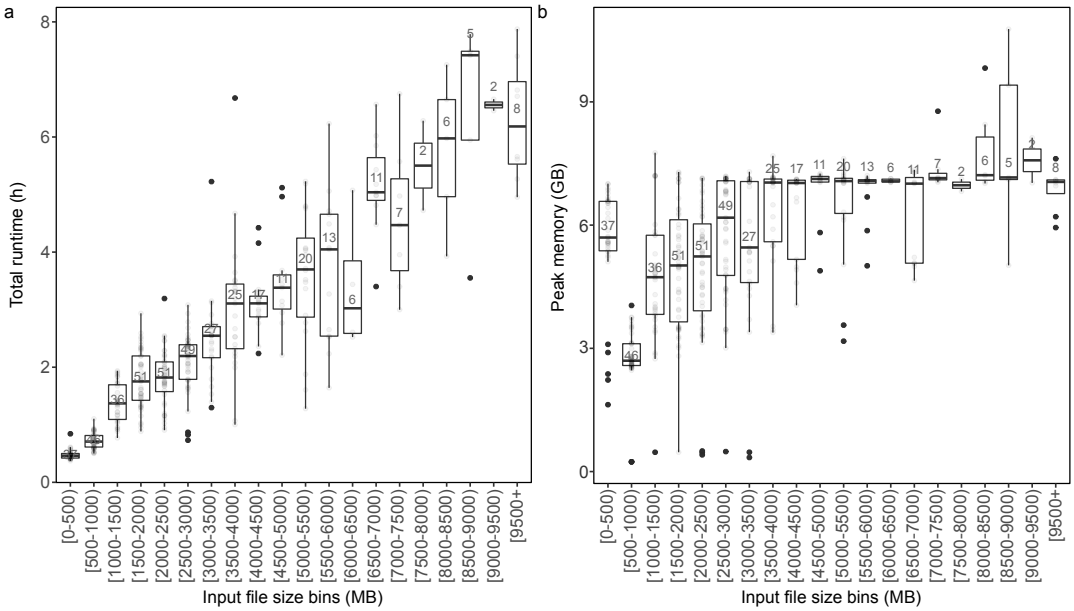

Figure 4.6: **PEPATAC is computational efficient**. (a) Pipeline runtime scales linearly with input file size. (b) Pipeline memory use peaks between 5-9GB.

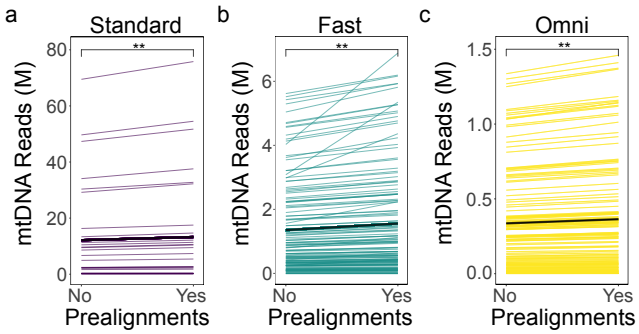

Figure 4.7: **Prealignment increases mtDNA alignment.** Within Standard (a), Fast (b), and Omni (c) ATAC-seq library preparation protocols, every sample shows increased mtDNA alignment when utilizing prealignments (The gray lines represent the mean increase within each protocol. ** = $p < 0.001$; t-test (mu = 0) with Benjamini-Hochberg correction.)

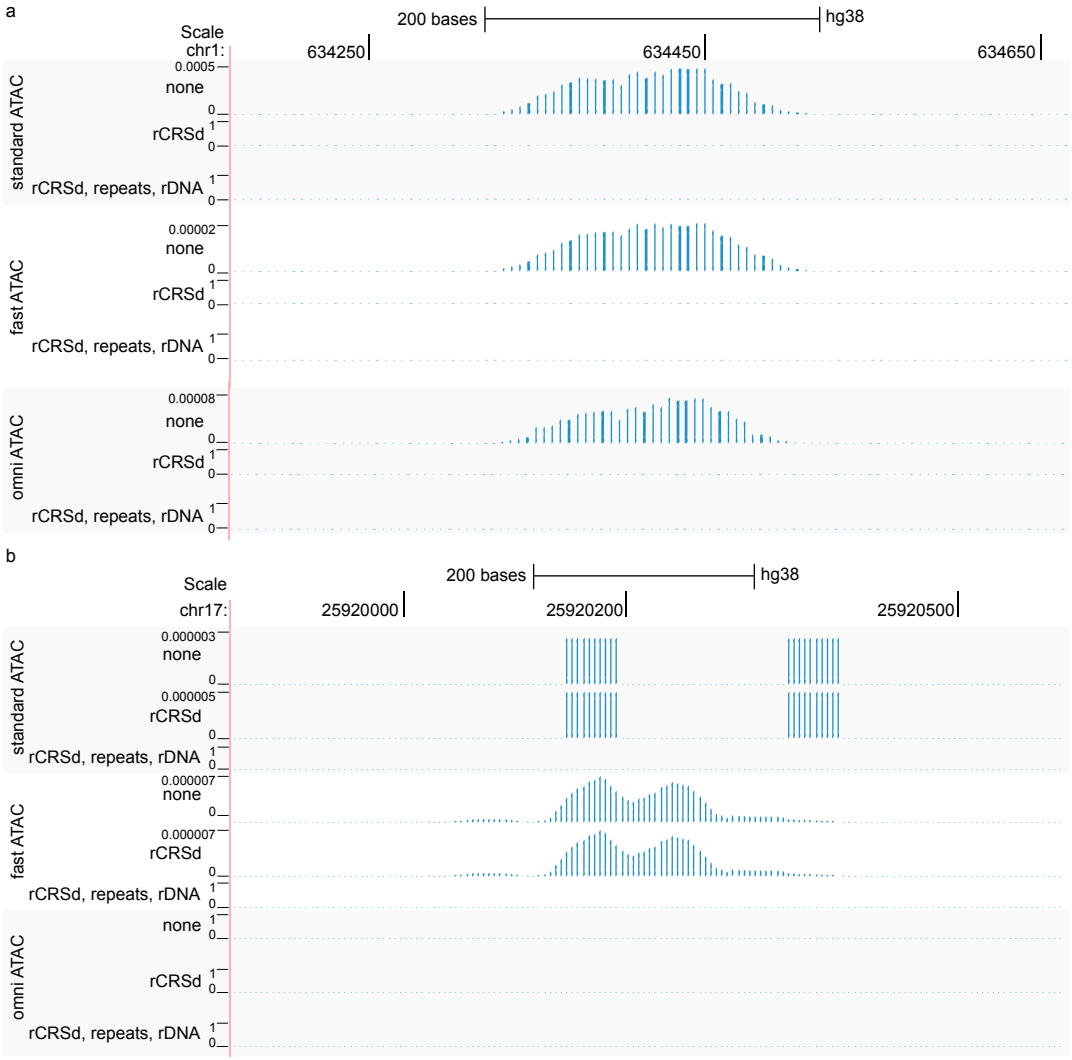

Figure 4.8: **Prealignment (and improved ATAC-seq library preparation protocols) successfully deplete signal from NuMTs, repeat regions, and high signal regions**. (a) Even where improved library preparation protocol leads to a NuMT annotated peak, prealignment successfully removes the spurious signal. (b) Both omni ATAC and prealignment to mitochondria *and* repeats and ribosomal sequence successfully depletes a spurious signal.

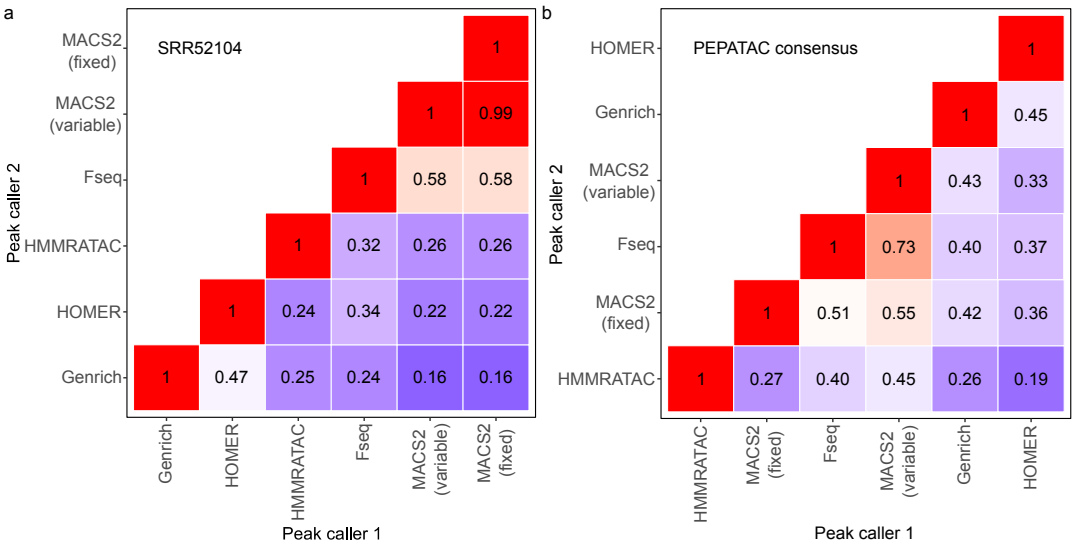

Figure 4.9: **Peaks are comparatively dissimilar between the five optional peak callers**. (a) For a single sample, MACS2 derived peaks, both with fixed and variable width peaks, are the most similar to Fseq called peaks. Genrich and HMMRATAC are the most unique among peak callers. (b) After PEPATAC consensus peak generation, HMMRATAC becomes even more dissimilar from the results derived from alternative peak callers.

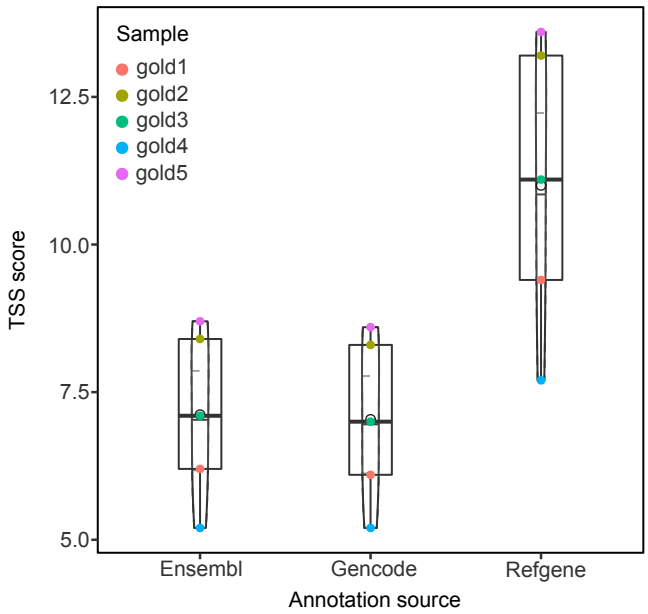

Figure 4.10: **The TSS enrichment score is dependent on the annotation source**. Refgene TSS annotations, which include the predominant TSS annotation only, produces the highest TSS enrichment score.

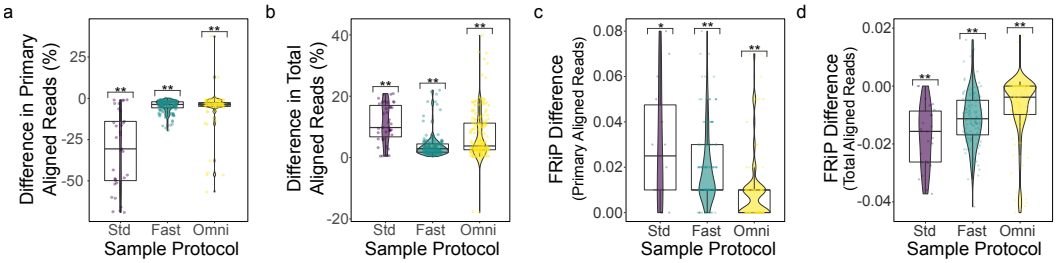

Figure 4.11: **Prealignment changes the relationship between primary genome and total aligned reads and the fraction of reads in peaks (FRiP) is dependent on mapping strategy.** (a) The number of primary, nuclear genome mapped reads is reduced when using prealignments. (b) However, the total number of mapped reads is increased with prealignments due to more specific read mapping. (c) The FRiP is increased with prealignments when using primary, nuclear genome mapped reads as the denominator. (d) In contrast, when using the total mapped reads the FRiP is reduced when using prealignments due to a larger mapped read pool in the denominator (* = $p < 0.01$; ** = $p < 0.001$; t-test (mu = 0) with Benjamini-Hochberg correction).

4.5.2 Supplemental_file_1

Supplemental_file_1.csv is the PEP-formatted sample table for the primary dataset. Samples are defined by protocol, whether standard, fast, or omni, and include accession numbers for access through the Gene Expression Omnibus [374].

- Supplemental_file_1.csv.

4.5.3 Supplemental_file_2

Supplemental_file_2.xlsx contains two sheets. The "jaccard_similarities" sheet includes tables representing the results of `bedtools intersect` between each independent peak caller software for 1) the `PEPATAC` derived consensus peak set, and 2) for an individual sample (SRR5210416) between each peak caller. This sheet also includes the average jaccard statistic for each peak caller. The "blacklisted_regions" sheet compares the number of peaks generated by each peak caller that overlap blacklisted regions [358].

- Supplemental_file_2.xlsx.

4.5.4 Supplemental_file_3

Supplemental_file_3.xlsx includes three sheets for a standard ATAC (SRR5427804), fast ATAC (SRR2920492), and omni ATAC (SRR5427806) sample that has been run through `PEPATAC` with 1) no prealignments, 2) mitochondrial prealignment (rCRSd: the revised Cambridge Reference Sequence doubled genome), and 3) mitochondrial, human repeats, and rDNA prealignments. In each sheet, for the highest scoring peaks, individual peak fasta sequences (included) were aligned with `BLAST` [372] and top scoring annotations recorded. If the peak overlaps a known blacklisted region [358], this is also marked.

- Supplemental_file_3.xlsx.

4.5.5 Supplemental_file_4

Supplemental_file_4.csv is the PEP-formatted sample table for the performance testing dataset. Accession numbers for file access through the Gene Expression Omnibus [374] are included for each sample.

- Supplemental_file_4.csv.

5 MEF2 family of transcription factors contribute to renin cell identity

Following the review and development of tools devoted to chromatin-related or chromatin-based assays, I next sought to apply this knowledge towards a specific biological question. In collaboration with Dr. Alexandre Martini and the University of Virginia Pediatric Center of Excellence in Nephrology, we sought to improve the field's understanding of the renin cell phenotype by investigating the chromatin landscape throughout renin cell development. Dr. Martini conceived the initial experiment, isolated mouse kidney cells, and prepared libraries for single-cell ATAC-seq and RNA-seq. I performed subsequent analysis and integration of the single-cell ATAC-seq and RNA-seq libraries. Here, I present the results of this analysis and report novel findings on the significance of the MEF2 family of transcriptions factors for renin cell development.

5.1 Background

By performing independently paired scATAC-seq and scRNA-seq at four developmental time points (E12, E18, P5, and P30) during mouse kidney development, we sought to identify epigenetic markers of renin cell identify and to construct a developmental trajectory from early Foxd1+ progenitors to mature renin cells. In effect, we have constructed a single-cell atlas of chromatin accessibility along a comprehensive time course of renin cell development.

Renin secreting cells are restricted in adult mammals along the walls of renal arterioles near the entrance to the glomeruli, and are therefore known as juxtaglomerular (JG) cells (Fig. 5.1a) [228–230]. These juxtaglomerular (JG) cells are critical for survival through the maintenance of homeostasis via the release of the hormone-enzyme renin in response to minute changes in blood pressure [226–228]. Renin release initiates a cascade that produces angiotensin II, leading to vasoconstriction and blood pressure increase. This renin-angiotensin system (RAS) is a key factor in cardiovascular pathologies [375–380]. Even in early hypertension, sustained activation of RAS signaling promotes vascular hypertrophy and dysfunction [381, 382]. RAS signaling is also relevant to diabetes, chronic kidney disease, dementia, and numerous cancers [383–388]. The standard therapy for hypertensive disorders and chronic kidney disease is the use of anti-RAS inhibitors and blockers [230, 389–391]. Unfortunately, inhibition of RAS leads to chronic production of renin and severe kidney disease [382, 392–394]. From the perspective of human health and disease, a better understanding of the control and formation of renin-expressing cells is essential to address chronic effects of RAS inhibition. Not only do renin cells play a vital role in homeostasis directly, they are also progenitors for multiple cell

types that retain the memory of the renin phenotype and are able to restore this phenotype to produce renin under stress [228, 230]. Despite the clear importance of these cells for organism health, we still do not fully understand their development.

Major efforts to elucidate determinants of renin cell identity have uncovered a number of important pathways and genomic regions integral to renin-expressing cells. The cAMP pathway has been shown to stimulate renin gene transcription and subsequent release (Fig. 1.3) [236–238]. The renin gene contains a cAMP responsive element where the histone acetyl transferases CBP/p300 can bind to regulate renin expression [239–241]. Additionally, the final common effector of the Notch signaling pathway, RBP-J, is necessary to maintain renin expression and modulates the plasticity of SMCs and mesangial cells to restore renin expression [230, 242–244]. RBP-J also regulates Akr1b7 which is co-expressed with renin and serves as an additional marker of mature renin cells [230, 245]. Understanding the epigenetic changes that occur to regulate the renin phenotype is on-going. Past work in our group identified a set of super-enhancers unique to renin cells [228]. The primary super-enhancer was found just upstream of the renin gene (Ren1) and is thought to be responsible for the restoration of renin phenotype in renin cell descendants [228]. Despite this knowledge of renin control, we are only beginning to uncover the epigenetic changes that occur along the differentiation trajectory of renin cells. An improved understanding of the dynamic genetic and epigenetic changes that occur in renin differentiation is necessary to better understand kidney pathologies and the effects of therapeutic targeting in cardiovascular disease.

While past efforts have greatly contributed to our knowledge or promoter and enhancer elements affecting renin expression [235, 238, 241, 243, 391, 395–397], no comprehensive study has been performed to delineate the individual factors that contribute to renin cell development in animals. Based on our own and others' past work, we sought to define the epigenetic changes and factors that govern the identity and plasticity of renin-expressing JG cells. Here, we produce a single-cell atlas of open chromatin and gene expression from progenitors to mature juxtaglomerular cells in the developing kidney and identify the MEF2 family of transcription factors as important contributors to JG cell formation and function.

5.2 Results

We report a better understanding of the underlying epigenetic changes that occur during the formation of renin-expressing (JG) cells in the kidney. Although we lack specific spatial positioning of our cell clusters, we term the highly accessible and highly expressing renin cells as juxtaglomerular cells to be clear on the putative identify of this cell population. We

identify JG cells using markers of these cells present in our cell subpopulations (i.e. identify JG cell marker gene expression and accessibility). This relies on the incorporation of known markers of JG cells we and others have previously identified (see Methods for complete details) [235, 391, 398]. Because we begin by isolating Foxd1+ cells, Foxd1 serves as a marker of cells ultimately forming the JG population. Furthermore, JG cells are predominantly identified by the expression of renin. Therefore, we looked for cells which express JG gene markers and have open chromatin at Ren1.

5.2.1 Overview of the epigenetic landscape of juxtaglomerular cell development

To determine the identity of the mature renin-expressing JG cells in the mouse kidney, we extracted mouse kidneys and isolated single cells across four developmental time points, E12, E18, P5, and P30 (Fig. 5.1a). At E18, Foxd1+ cap mesenchyme cells are the progenitors of renin producing cells, vascular smooth muscle cells (VSMCs), mesangial cells, and pericytes (PCs) found in the fully formed mature kidney by P30. By P30, the renin cells are confined to the juxtaglomerular region, yet cells in this lineage retain the ability to revert to renin-expressing phenotypes (Fig. 5.1a). In this lineage-tracing model, cells that express Foxd1 at any point during development are marked by expression of GFP (Fig. 5.1b). Since the primordial metanephros is very delicate and small at E12, we identify GFP+ animals by lung "squashes" and proceed through a different cell isolation protocol (See Isolation of kidney single cells: E12) without posterior FACS. For the three later timepoints, we leverage GFP expression to FACS sort single cells isolated at each time point that are positive for GFP (See Isolation of kidney single cells: E18, P5 and P30). All isolated cells are subsequently subjected to independent scATAC-seq and scRNA-seq experiments (Fig. 5.1b). We mapped the transposase-accessible chromatin and gene expression at the single-cell level using the 10X Genomics Chromium platform, and integrated these datasets to perform a combinatorial analysis of the transcriptome and accessibilome of Foxd1 lineage cells (Fig. 5.1b). The scATAC-seq samples formed clusters predominantly separated by developmental time point with P30 cells showing the most spatially removed clustering profile (Fig. 5.1c). Independent scRNA-seq cells were clustered and revealed 21 distinct annotated clusters (Fig. 5.1d, see scRNA-seq cell clustering and scRNA-seq cell identification). We integrated the scRNA-seq data with the scATAC-seq and performed label-transfer to ultimately annotate 23 open chromatin derived clusters (Fig. 5.1g, See Integrating transciptome and accessibilome).

To identify the subset of cells representing mature JG cells, we looked for canonical markers of JG cell identity, the genes Ren1 and Akr1b7 (See Identifying renin cells)[391]. By evaluating these marker genes from the gene activity scores (Fig. 5.1e) and integrated gene expression

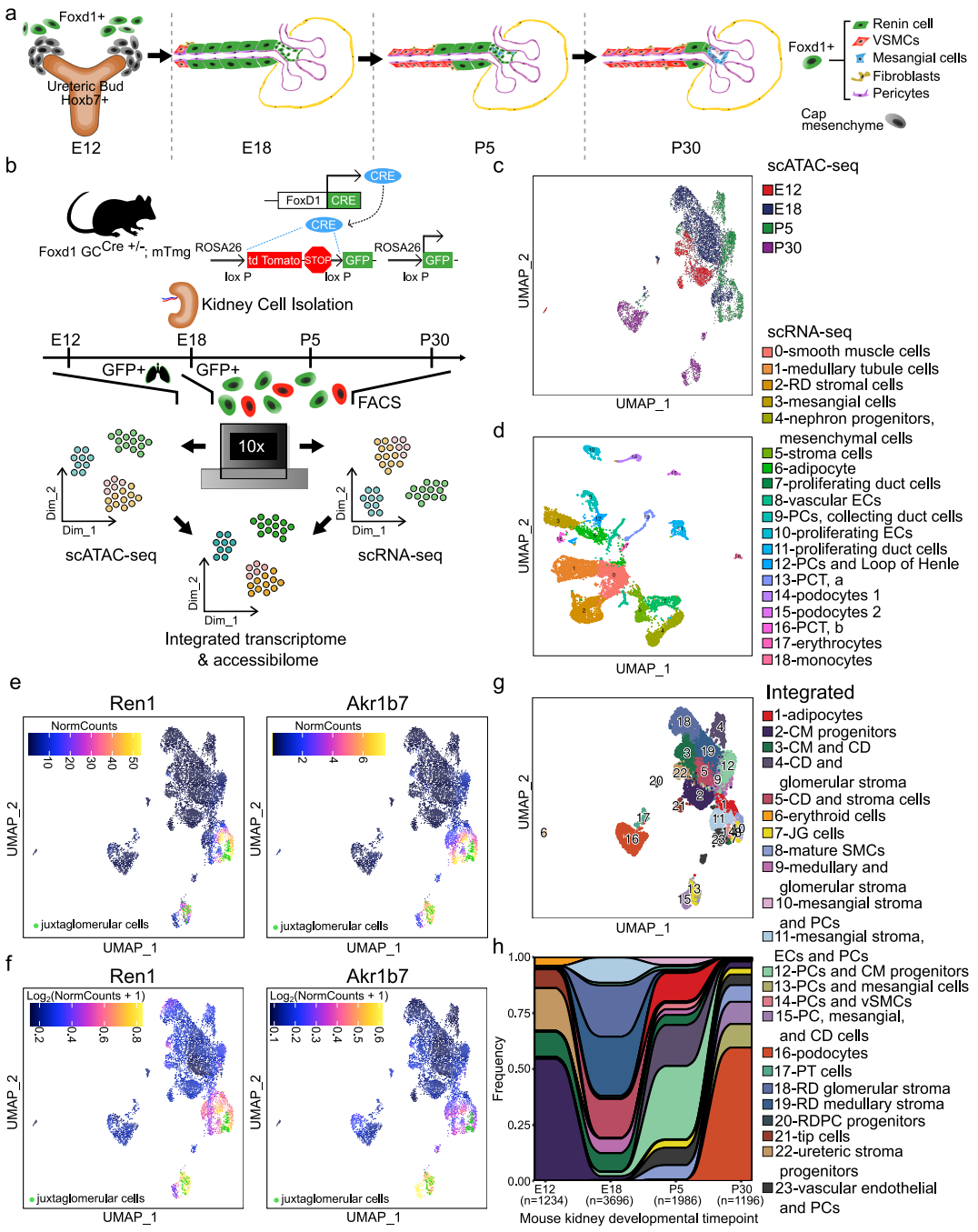

Figure 5.1: **Overview of the experimental design to identify the renin cell developmental trajectory** . (a) Foxd1 progenitors in the cap mesenchyme in early E12 differentiate through E18, P5, and P30 to lead to mature renin expressing cells in the juxtaglomerular region. (b) Kidneys isolated from Foxd1-CRE recombinase mouse lineage-tracing model are sorted on Foxd1-derived GFP expression and single-cell ATAC-seq and RNA-seq is performed. (c) UMAP visualization of scATAC-seq data separated by time point. (d) UMAP visualization of scRNA-seq data with annotated cell clusters. (e) UMAP visualization of gene activity scores for canonical JG markers Ren1 and Akr1b7. (f) UMAP visualization of gene expression for canonical JG markers Ren1 and Akr1b7. (g) UMAP visualization of integrated scATAC-seq and scRNA-seq data with annotated cell clusters. (h) Cell frequency distribution across developmental time point. Numbers below timepoints represent total number of single cells at each time point. CM: cap mesenchyme; CD: collecting duct; JG: juxtaglomerular; PC: pericyte; EC: endothelial cell; SMC: smooth muscle cell; PT: proximal tubule; PCT: proximal convoluted tubule; RD: rapidly dividing.

(Fig. 5.1f), we identified a subpopulation of cells representing the likely juxtaglomerular population. We also explored the distribution of annotated cells in the open chromatin data across the kidney developmental time points to evaluate lineage contributions to the JG cells (Fig. 5.1h).

5.2.2 Differentiation trajectory of juxtaglomerular cells

Because much remains unknown about the epigenetic changes that lead to JG cell identity, we next sought to identify the regions of open chromatin, corresponding transcription factors and their linked gene score and expression, to investigate what genetic and epigenetic features define JG cells. Using previously identified genetic markers of renin-expressing juxtaglomerular cells (See Identifying renin cells) we defined a pseudo-time trajectory of cells with high Foxd1 expression and accessibility early leading to cells with high expression and accessibility of Ren1 and Akr1b7 (Fig. 5.2a, Fig. S**??**).

By looking within peaks that mark individual cell clusters along the developmental trajectory, we can identify enriched motifs that most define individual cell types (Fig. 5.2b; See Motif annotations and enrichment). Here we identify very high enrichment of the myocyte enhancer factor 2 (MEF2) family of TFs as well as Nfix in both mature SMCs and JG cells (Fig. 5.2b). The nuclear factor I family of TFs (including Nfix, Nfic) have been previously reported to bind to renin promoter and enhancer regions in the genome [396, 397], and we confirm motif occurrences of these and other factors at Ren1 (Fig. S5.7, Fig. S5.8). The motifs for Rfx2, Ebf1, Bach2, Smarcc1 are also enriched in both of these late time point populations (Fig. 5.2b). When distinguishing mature SMCs and JG cells, we find a reduced enrichment for Grhl1, Snai2 and Smad5 specifically in JG cells (Fig. 5.2b).

We also utilized our integrated pseudo-time analysis to identify positive drivers of differentiation along the trajectory (See Positive transcription factor regulators, Fig. S5.11). Here we link gene scores or gene expression to their corresponding motifs and uncover a number of genes and motifs linked across pseudo-time and both gene score (Fig. 5.2c) and expression (Fig. 5.2d). Among the cell populations containing JG cells, the MEF2 family, Ets1, Ebf1, Junb, and Stat3 are enriched across both integrative approaches (Fig. 5.2c,d).

We also identify enriched gene scores, gene expression, regulatory regions, and TF motifs across pseudo-time that are not exclusively positively correlated. By the emergence of mature JG cells (late P5 to P30), there are enriched gene scores for Foxs1, Zfp36, Junb, Fosl2 and Dusp1 (Fig. S5.9a).

Foxs1 has been previously implicated as containing SNPs contributing to high blood-pressure

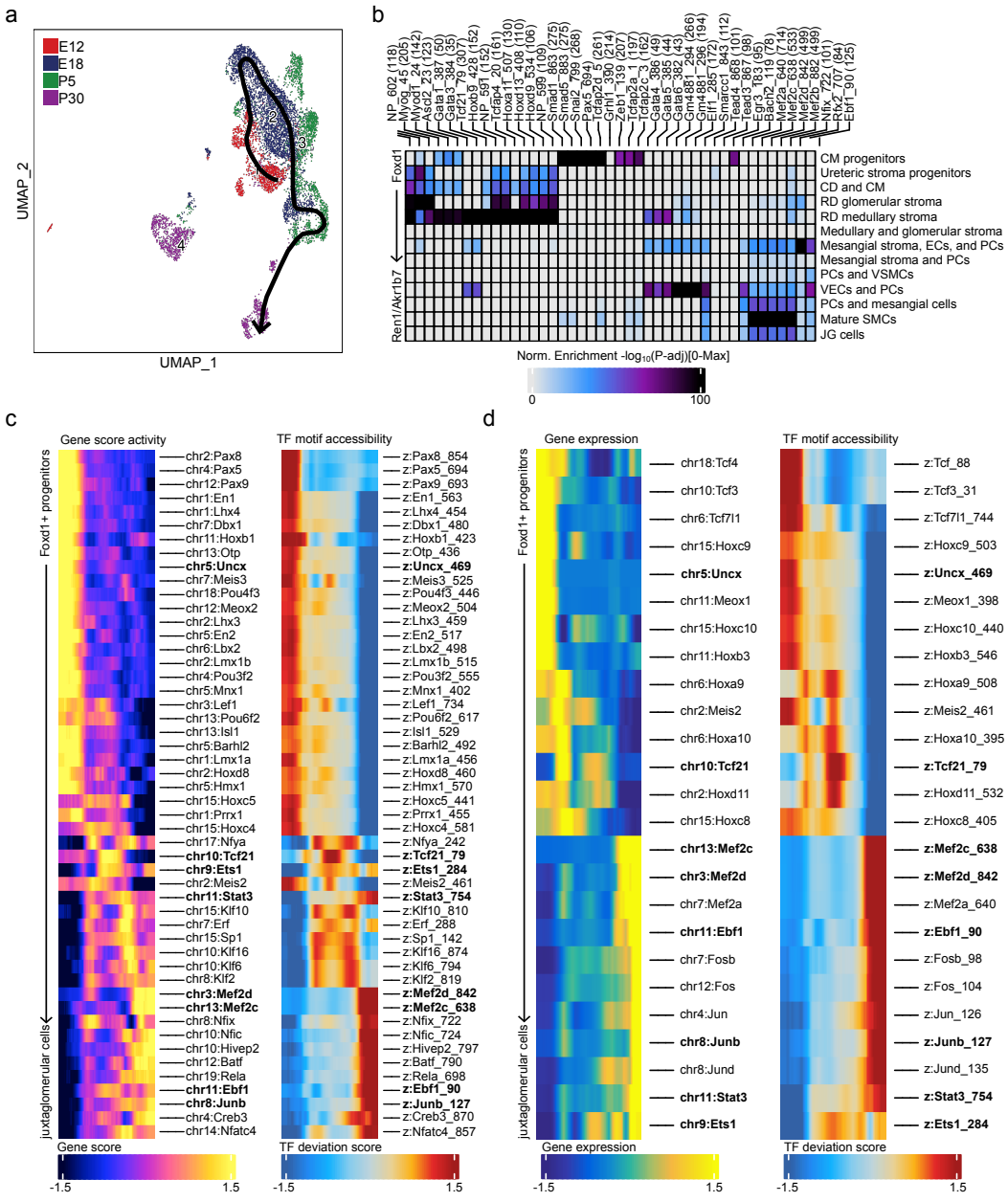

Figure 5.2: **Epigenomic differentiation trajectory uncovers renin cells by post-natal day 5 in mouse kidney development**. (a) UMAP visualization showing the pseudo-time trajectory across developmental time points. Arrow head represents the end point of the trajectory. (b) Heatmap of motif hypergeometric enrichment-adjusted P values within the marker peaks of each JG trajectory cluster. Color indicates the motif enrichment (-log10(P value)) based on the hypergeometric test. (c) Integrated pseudo-time analysis of positively correlated gene scores and corresponding motifs. (d) Integrated pseudo-time analysis of positively correlated gene expression and corresponding motifs. Bold text indicates genes and motifs identified in both integrated approaches. CM: cap mesenchyme; CD: collecting duct; EC: endothelial cell; JG: juxtaglomerular; PC: pericyte; RD: rapidly dividing; SMC: smooth muscle cell; VEC: vascular endothelial cell; VSMC: vascular smooth muscle cell

indicative of its importance to RAS signaling [399]. As Foxd1 is already an established marker of renin-expressing cells in the kidney, Foxsl may contribute as a secondary marker of this trajectory. Additionally, the forkhead box family members play a role in regulating the Wnt signaling pathway along with Tcf and Lef transcription factors [380, 400–402], both of the latter of which are positive TF regulators in our pseudo-time trajectory (Fig. S5.11). Together, these findings confirm a role for Wnt signaling in renin-cell development [402, 403].

Fosl2 has been previously identified to regulate TGF-β1 [404], and TGF-β1 is itself induced by renin expression [405, 406] suggesting a possible novel target in individuals with chronic stimulation of the renin-angiotensin system, in addition to a possible mechanism of renin-expressing cell dysfunction when treating with RAS inhibitors.

We also identify enrichment of the expected marker gene Ren1 and mature smooth muscle markers Acta2, Tagln, and Crip1 (Fig. S5.9b). Late in pseudo-time where JG cells emerge, there are enriched regulatory regions in chromosomes 1, 2, 4, 6, 7, 8, and 12 (Fig. S5.9c) indicating possible trajectory defining regulatory regions. Looking closer, we then annotated all marker regulatory regions between clusters along the JG differentiation trajectory to look for enrichment of known functional classes between clusters (Fig. S5.10). Overall, we again identify enrichment of motifs for the MEF2 family of transcription factors, as well as for Smarcc1, Bach1 and Bach2, Nfix, Fos, and Jun (Fig. S5.9d).

5.2.3 Transcription factors contributing to juxtaglomerular cell development

Next we investigated enriched motifs and their corresponding expression and activity scores for their potential contribution to JG differentiation. First we performed footprinting (See TF footprinting) to look for enrichment of late acting TFs and confirmed overall enrichment for Smarcc1 (Fig. 5.3a), Nfix (Fig. 5.3b), Mef2c (Fig. 5.3c), and Bach2 (Fig. 5.3d) in the late differentiation clusters which include the JG cell subpopulation. A pattern emerges whereby different TFs with aggregate enrichment of their footprints display differential patterns of activity. We categorized these and additional TFs as enriched early, middle, or late during differentiation (Fig. S5.12).

Of these initially identified TFs (see Motif annotations and enrichment), Smarcc1 gene activity scores (Fig. 5.3e) and expression (Fig. 5.3i) peak early to middle along the JG differentiation trajectory before dropping to near zero at terminal JG differentiation. Smarcc1 is part of the SWI/SNF chromatin remodeling complex and can interact with a number of different transcription factors [407, 408]. Data is suggestive for a role of Smarcc1 in early remodeling of regulatory regions critical for JG cell identity.

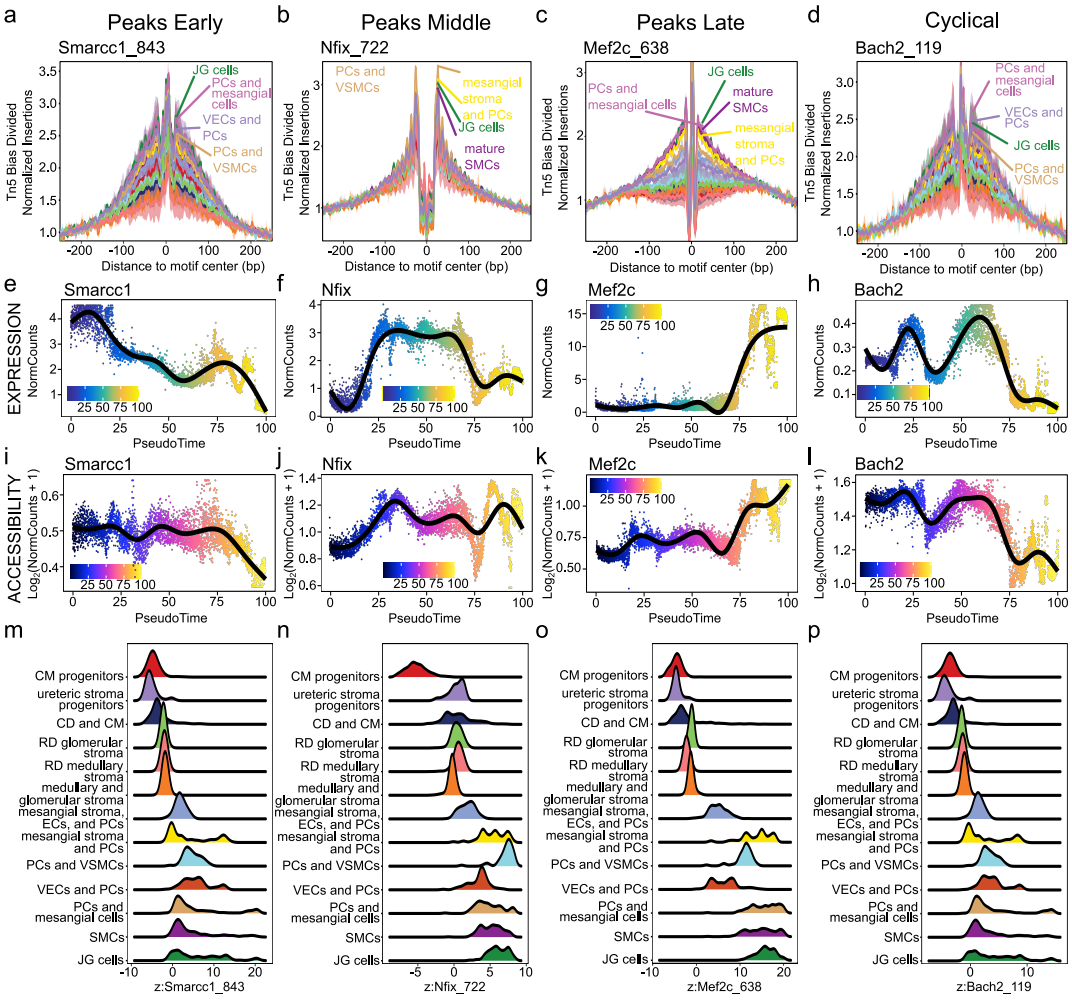

Figure 5.3: **TF expression, accessibility, and enrichment uncover patterns of differentiation**. Footprints for Smarcc1 (a), Nfix (b), Mef2c (c), and Bach2 (d) are enriched in the JG cluster and parent clusters of JG and SMC cells. The expression pattern of enriched TFs in the JG cluster illustrates early (e), middle (f), late (g), and cyclical (h) patterns of transcript abundance. Gene score activity recapitulates expression patterns of early (e), middle (f), late (g), and cyclical (h) activity. TF deviation scores highlight enrichment of Smarcc1 (m), Nfix (n), Mef2c (o), and Bach2 (p) in late differentiation along the JG trajectory. CM: cap mesenchyme; CD: collecting duct; EC: endothelial cell; JG: juxtaglomerular; PC: pericyte; RD: rapidly dividing; SMC: smooth muscle cell; VEC: vascular endothelial cell; VSMC: vascular smooth muscle cell

Nfix gene activity is maximal during the middle period of differentiation (Fig. 5.3f), with gene expression peaking before dropping and rebounding at terminal JG formation (Fig. 5.3j). As a known factor binding to the renin promoter and enhancer regions [396, 397], we provide novel evidence suggesting a time-dependent mechanism of action for nuclear factor I family members.

Mef2c gene activity (Fig. 5.3g) and expression ((Fig. 5.3k) both peak during the initial differentiation into late time point clusters including pericytes and mesangial cells, before peaking again late in the clusters containing SMCs and JG cells. The MEF2 family of TFs includes four proteins, Mef2a/b/c/d, and each play important roles in cardiac and skeletal muscle tissues where they interact with chromatin remodeling factors and other transcriptional regulators [409–413]. Additionally, MEF2 target gene activation has been directly linked to stimulation by p300 [414, 415], which is itself critical to remodeling of chromatin at the renin locus [228]. Despite the similarity of target motifs for MEF2 members, previous work has demonstrated that individual MEF2 members regulate non-overlapping gene programs [412], suggesting distinct roles for MEF2 members in renin regulation previously unknown.

Interestingly, Bach2 undergoes a cyclical pattern of gene activity (Fig. 5.3h) and expression ((Fig. 5.3l) emphasizing a possible role in the cell cycle of differentiating JG cells [416, 417]. Bach2 is itself a transcriptional repressor which forms heterodimers with small Maf proteins and bind at Maf-recognition elements (MARE) of target genes [418, 419]. MAREs share strong sequence conservation with CRE elements [420] with a cAMP response element present at the renin locus and essential for renin expression [228, 230, 421]. It is possible then that Bach2 interacts with companion factors to repress expression of genes that direct progenitors towards a JG cell fate as it does in other cells [422]. The cyclical expression and resultant reduction in Bach2 expression in late developmental cell populations supports such a role.

Finally, we evaluated the deviation in TF motif enrichments on a per cell basis to further confirm enrichment of these identified TFs as relevant to JG differentiation. Here we show that each of the above TF motifs is preferentially enriched in late forming cell clusters, including the JG population (Fig. 5.3m-p).

5.2.4 MEF2 family of TFs separates JG cells from mature SMCs

We then further narrow our focus by investigating what factors specifically differentiate the late emerging clusters of SMCs and JG cells. Only a few (24) accessible regions differentiate these two similar clusters (Fig. 5.4a), however when we investigate enriched motifs in SMCs (Fig. 5.4b) or in JG cells relative to SMCs (Fig. 5.4c) we see some familiar actors. In

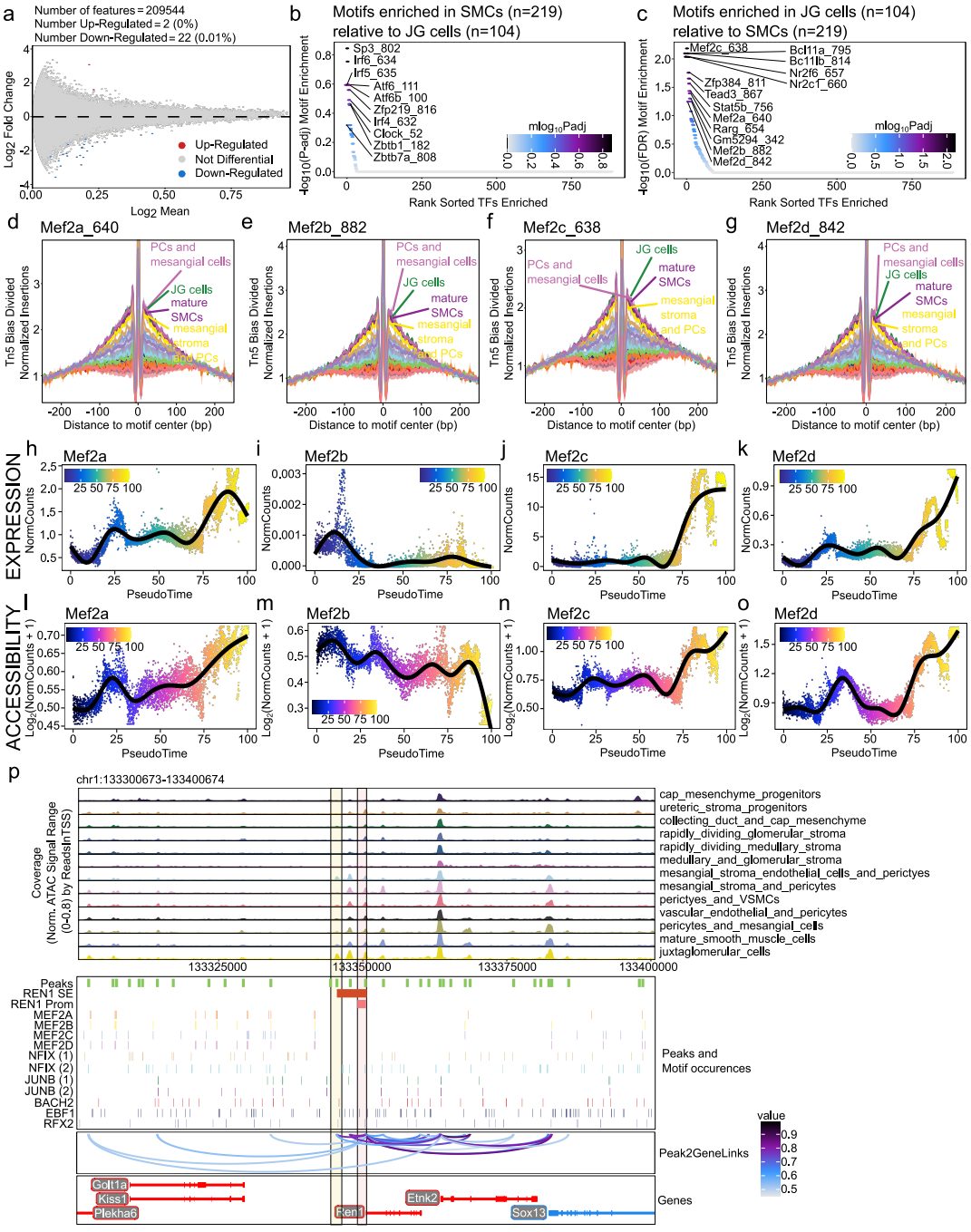

Figure 5.4: **MEF2 family of TFs uniquely defines the terminal JG population**. (a). MA plot of the differential regulatory regions between SMCs and JG cells. (b) Top enriched motifs in SMCs as compared to JG cells. (c) Top enriched motifs in JG cells compared to SMCs. Mef2a (d), Mef2b (e), Mef2c (f), and Mef2d (g) footprints are enriched in JG cells. Mef2a expression (h) and gene score activity (l) peaks just prior to terminal differentiation into JG cells. Mef2b expression peaks in early differentiation (i) with the corresponding gene score activity steady early to middle before plummeting at differentiation into JG cells (m). Mef2c (j,n) and Mef2d (k,o) expression and gene activity both peak at terminal JG differentiation. (p) Browser tracks identify preferentially enriched open chromatin in the JG cell cluster at Ren1. Motif occurrences of enriched TFs identify putative binding sites in open and co-accessible peaks. Yellow fill box highlights uniquely enriched peak in JG cluster. Pink fill box highlights JG promoter region. JG: juxtaglomerular; SE: super-enhancer; SMC: smooth muscle cell

particular, the entire MEF2 family is preferentially enriched in JG cells. We also looked at patterns of enrichment of each of the top enriched motifs that distinguish JG cells from SMCs (Fig. S5.13). The MEF2 (Mef2a (Fig. S5.13a), Mef2c (Fig. S5.13b)) family of TFs, Zfp384 (Fig. S5.13c), Stat5b (Fig. S5.13e), the nuclear receptor subfamily 2 members (Nr2f6 (Fig. S5.13h), Nr2c1 (Fig. S5.13i)), and Rarg (Fig. S5.13j) are universally enriched in late differentiation and within the JG cell population.

With the data purporting to show a strong role for MEF2 family members contributing to JG cell differentiation, we next looked for overall enrichment of MEF2 family footprints along the differentiation trajectory and identify each family member as enriched in JG cells (Fig. 5.4d-g). Individual MEF2 members show different activities when we compare gene activity scores and gene expression for each independent MEF2 member. Mef2a gene activity and expression peak in late differentiation (Fig. 5.4h,l). Conversely, Mef2b spikes early (Fig. 5.4i) and experiences a precipitous drop (Fig. 5.4n) in expression in late differentiation. Both Mef2c and Mef2d display overall similar patterns of activity, spiking briefly mid-to-late before showing maximum activity and expression in the terminally differentiated JG containing populations (Fig. 5.4k,o).

Finally, we looked at whether any of the enriched TFs so far identified have putative binding sites at Ren1 (See Peak co-accessibility and peak to gene links). The previously mentioned TFs Bach2, Ebf1, Nfix, and Rfx2 all contain motif occurrences in either the Ren1 promoter and/or super-enhancer regions (Fig. 5.4p). While we do not identify MEF2 family motif occurrences in either location, there are sites present in co-accessible regions with predominant peaks of open chromatin in the JG cell population (Fig. 5.4p). With past evidence suggesting a role of MEF2 family members in recruiting chromatin remodelers including p300, this supports chromosomal conformation as an important regulatory feature of JG cell identify in conjunction with the action of MEF2 TFs.

5.3 Discussion

Understanding the regulatory landscape of renin expressing cells is necessary to better understand the control and function of this rare cell population with critical roles in health and disease [229, 230, 235, 391, 398]. Our current understanding of the epigenomic regulatory landscape in renin expressing cells is limited, therefore it is urgent we expand our knowledge to understand consequences of disruptions to the renin-angiotensin system that occur in human health and disease and the effects of drugs targeting this pathway.

Here, we employed high-throughput 10X-based scRNA-seq and scATAC-seq technology

to simultaneously measure the transcriptome and accesibilome of progenitors and mature renin-expressing cells in the mouse kidney. Our integrated profiles revealed a sequential differentiation pattern from early kidney development to mature kidney. Following the first known identification of the renin-expressing population of juxtaglomerular cells in mouse kidney, we identified substantial differences in the TF regulators of each cell subpopulation along the differentiation trajectory.

In particular, the MEF2 family of transcription factors are enriched in late differentiation clusters, including the juxtaglomerular cell population. Previous studies of the importance of MEF2 members in angiogenesis [415], and the interaction of MEF2 members with p300 [411, 414] provide a foundation for future *in vivo* and *in vitro* experiments to directly interrogate the individual roles of MEF2 members. Overall, we have provided the first identification of juxtaglomerular cells in a single-cell atlas of kidney open chromatin, and highlighted a number of novel factors important for their differentiation and function.

5.4 Materials and Methods

5.4.1 Mouse model (from the UVA Pediatric Center of Excellence in Nephrology)

All animals were maintained in a room with controlled temperature and humidity under a 12-hour light/dark cycle. All animals were handled in accordance with the National Institutes of Health guidelines for the care and use of experimental animals, and the study was approved by the Institutional Animal Care and Use Committee of the University of Virginia.

To generate *Foxd1 cre/+;R26R-mTmG* mice, we crossed *Foxd1 cre/+* mice (16, 24) with the *R26R-mTmG* mice [423, 424]. scRNA-seq and scATAC-Seq were performed using the 10X Genomics technology according to the respective protocol [425]. For the scATAC-Seq, we performed the nuclei isolation following the manufacturer protocol [426], and all the experiments targeted at least 2000 nuclei, while the scRNA-seq always targeted more than 1000 cells. The experiments were performed at specific time points of the mouse kidney development: E12 (embryonic day 12), E18 (embryonic day 18), P5 (five days old) and P30 (one month old). Please, note that mice nephrogenesis and vascular development starts at E11.5 and continues after birth for about 3-7 days, respectively [233].

5.4.2 Isolation of kidney cells

Isolation and subsequent library preparation of kidney cells for single-cell ATAC-seq and RNA-seq was performed by Dr. Alexandre Martini.

5.4.2.1 Isolation of kidney single cells: E12 Pregnant mice were injected with tribromoethanol at enough dose to keep alive but anesthetized. Then, the small fetuses were removed one by one. A small squash from the lung was used to identify the GFP pups, using an EvosFLC cell imaging system (LifeTechnologiesTM, California, USA). Once they were identified, the metanephros area was removed under microscopy and harvested in ice cold dPBS. The tissue was then minced carefully with a razor blade and transferred to 1.7mL Eppendorf tube. 300 μL of TryPLETM Express (Gibco, New York, USA) was added and incubated at 37°C for 5 minutes. Then, 600 μL of DMEM + 5% FBS was added to the Eppendorf tube. The mixture was homogenized by pipetting up and down. The solution was then filtered with a 40 μm nylon cell strainer. Tissue chunks were removed from the mesh surface, placed again in an Eppendorf tube and the process repeated a second time. Meanwhile, the flow-through filtrate was placed in a new Eppendorf tube and centrifuged at 4°C, 150g for 5 minutes. The supernatant was removed and the pellet resuspended with resuspension buffer. All the tubes were combined in the end and the cells were ready for single-cell capture.

5.4.2.2 Isolation of kidney single cells: E18, P5 and P30 Animals were anesthetized with tribromoethanol (300 mg/kg). P5 and P30 mice kidneys were excised and decapsulated. Then, the kidney cortices were dissected, minced with a razor blade, and transferred into a 15 mL tube with 5 mL of enzymatic solution (0.3% collagenase A [Millipore-Sigma], 0.25% trypsin [Millipore-Sigma], and 0.0021% DNase I [Millipore-Sigma]). The tubes were placed flat inside a shaking incubator (80 RPM) for 15 minutes at 37°C. The solution was pipetted up/down 10 times with a sterile transfer pipette and allowed to settle for 2 minutes, and the supernatant was collected in a fresh tube on ice. The enzymatic solution was added to the 15 mL tube containing the remaining undigested cortices, and the digestion procedure was repeated a total of 3 times. The supernatants were pooled and centrifuged at 800 g for 4 minutes at 4°C using a Sorvall RT7 refrigerated centrifuge (Sorvall, Newtown, CT). The cell pellet was resuspended in fresh buffer 1 (130 mM NaCl, 5 mM KCl, 2 mM CaCl2, 10 mM glucose, 20 mM sucrose, 10 mM HEPES, pH 7.4), and the suspension was poured through a sterile 100 μm nylon cell strainer (Corning Inc., Corning, NY) and washed with buffer 1. The flow-through was poured through a sterile 40 μm nylon cell strainer (Corning Inc.) and washed with buffer 1. The flow-through was centrifuged at 1,100 g for 4 min at 4°C. The cell pellet was resuspended in 1.5 mL of resuspension buffer [PBS, 1% FBS, 1 mM EDTA, DNAase I (Millipore-Sigma)]. DAPI (Millipore-Sigma) was added to the cells to identify the living cells. The GFP positive cells were collected by Fluorescent-Activated Cell Sorting (FACS)

[230] and resuspended in DMEM (Dulbecco's Modified Eagle Medium, Gibco, Netherlands) with 10% FBS (Fetal Bovine Serum) for immediate use. The sorters were either an Influx Cell Sorter (Becton Dickinson, Franklin Lakes, NJ) or a FACS Aria Fusion Cell Sorter (Becton Dickinson), both located at the Flow Cytometry Core Facility at the University of Virginia.

5.4.3 scATAC-seq library preparation

The initial steps of capture between scATAC-seq and scRNA-seq are similar. The GFP-positive cells are washed in dPBS with 0.04% BSA twice, and the sorted GFP-positive cells are washed in dPBS with 0.04% BSA twice. The cells are counted with the CellometerMini (Nexcelom Bioscience, Massachussets, USA). Sorted cells with viability higher than 80% and absence of clumps were chosen to proceed. However, the scATAC-Seq protocol requires nuclei isolation. Our experiments yielded less than 100,000 cells. Therefore, we have used the protocol designed for Low Cell Input Nuclei Isolation. Briefly, we centrifuge the cells at 500g at 4°C for 10 minutes, remove the supernatant carefully and resuspend the cells in a freshly prepared lysis buffer (1M Tris-HCl, 5M NaCl, 1M MgCl2, 10% BSA, 10% Tween-20). We have optimized the lysis incubation time for 7 minutes on ice. Next, we immediately added freshly made washing buffer (1M Tris-HCl, 5M NaCl, 1M MgCl2, 10% Tween-20, 10% Nonidet P40, 5% digitonin and 10% BSA). This washing is performed twice, and then supernatant is removed for a final wash with diluted nuclei buffer (nuclei buffer 20X, 10X Genomics, diluted in nuclease free water to 1X). Once this is complete, we again remove the supernatant, resuspend the nuclei pellet and count it with a Neubauer Chamber (Spencer, Buffalo, USA). In all our experiments we were able to target between 1000-5000 nuclei. Nuclei were loaded and captured with the Chromium System (10X Genomics, Pleasanton, CA) following the manufacturer's recommendation [425] using the Chromium Next GEM Chip H with reagents of Chromium Next GEM Reagent Kits v1.1 (10X Genomics). Initially, nuclei were incubated with the transposition mix, that includes a transposase, and later the GEMs are barcoded and PCR amplified to generate the cDNA. The cDNA is then cleaned with Dynabeads and SPRIselect, end-repaired, adaptor-ligated and amplificated by PCR. The constructed libraries were sequenced on an Illumina HiSeq 2500/4000 platform (150-bp paired-end reads).

5.4.4 scRNA-seq library preparation

Sorted GFP-positive cells were spun down at 500g for 10 minutes in a Sorvall RT7 refrigerated 4°C centrifuge (Sorvall, Newtown, CT). Then, the supernatant was carefully removed, and the cell pellet was resuspended in dPBS (Dulbecco's Phosphate Buffered Saline, Gibco, UK) with 0.04% BSA (UltraPureTM Bovine Serum Albumin,Invitrogen, Lithuania) 10 times

with wide-bore tips. The process was repeated once, and the cells were counted with the CellometerMini (Nexcelon Bioscience, Massachussets, USA). Experiments with more than 500 cells/μL, viability higher than 85%, and absence of clumps were chosen to proceed. Single cells were loaded and captured with the Chromium System (10X Genomics, Pleasanton, CA) following the manufacturer's recommendation [425] using the Chromium Next GEM Chip G with reagents of Chromium Next GEM Single Cell 3′ Reagent Kits v3.1 (10X Genomics). The Cell-Gel Beads in Emulsion (GEMs) were generated and incubated to generate the barcoded cDNA. cDNA was cleaned with Dynabeads and amplificated by PCR. cDNA was then enzymatically fragmented, end-repaired, poly-A tailed, adaptor-ligated, and amplificated by PCR. The constructed libraries were sequenced on an Illumina HiSeq 2500/4000 platform (150-bp paired-end reads).

5.4.5 Read preprocessing

5.4.5.1 Genome and transcriptome annotations We performed all downstream analyses using the mm10 genome. For Cell Ranger pipeline analyses, we used the `refdata-cellranger-arc-mm10-2020-A-2.0.0` reference data and for R-based analysis, the `BSgenome` package `BSgenome.Mmusculus.UCSC.mm10` was used.

5.4.5.2 scATAC-seq alignment and fragment matrix generation We processed fastq files from the 10X Genomics Single Cell ATAC platform using the Cell Ranger ATAC pipeline (version: `cellranger-atac-2.0.0`). Cell Ranger trims primer sequences using a modified `cutadapt` tool [319] before alignment using a modified `bwa-mem` algorithm [[365]; Li2010a]. Duplicate reads were removed based on the start, end, and hashed barcode of aligned reads. Finally, read fragments were corrected for Tn5 transposase binding biases and stored as position-sorted fragment files.

5.4.5.3 scATAC-seq quality control We processed Cell Ranger fragment files using the `ArchR v1.0.1` R package [224]. Arrow files were produced from each time point's fragments file. We determined the presence of putative cell doublets using the ArchR function `addDoubletScores`, which determines doublet identity by embedding synthetically created pseudo-doublets from the input data into the shared sample space. We removed cells that behaved like pseudo-doublets (n=599) prior to downstream analysis (Fig. 5.5a-d). We further excluded cells with TSS enrichment scores less than 10 as these low TSS scores are indicative of low signal-to-noise and poor quality (Fig. 5.5e-h) [427]. Finally, we evaluated fragment distributions for each developmental time point to confirm the expected periodicity of nucleosomal positioning (Fig. 5.5i-l).

5.4.5.4 scATAC-seq dimensionality reduction and batch correction Following quality control filtering, an iterative Latent Semantic Indexing dimensionality reduction [428, 429] technique using Singular Value Decomposition (SVD) to embed the most valuable sample information in low dimensional space [224] was applied (see ArchR `addIterativeLSI`). This reduced dimensionality matrix was further corrected for batch effects using the Harmony [430] algorithm originally developed for scRNA-seq data but extended to scATAC-seq in ArchR (using the function `addHarmony`) [224].

5.4.5.5 scATAC-seq cell clustering We then clustered the batch-corrected reduced dimensionality matrix to identify cells based on shared chromatin accessibility patterns using the `addClusters` function in ArchR [224]. These cluster embeddings were visualized using Uniform Manifold Approximation and Projection (UMAP) to evaluate dimensionality reduction and batch correction results across the integrated time points.

Because scATAC-seq data is inherently sparse, pseudo-bulk replicates were generated (`addGroupCoverages` function) by grouping similar single cells into aggregate profiles similar to bulk ATAC-seq data and then calling peaks. Peak calls were generated using `MACS2` [255] with the `addReproduciblePeakSet` function in ArchR.

5.4.5.6 scRNA-seq alignment and feature-barcode matrix generation Next, we processed scRNA-seq fastq files from the 10X Genomics Single Cell Gene Expression platform using the Cell Ranger pipeline (version: `cellranger-6.0.1`). The Cell Ranger pipeline trims non-template sequence and aligns reads using the STAR aligner to genomic and transcriptomic annotations [431]. Confidently mapped transcriptomic reads are grouped by barcode, UMI, and gene. The number of reads mapped to each gene was calculated using UMI-based counts. Finally, filtered UMIs were mapped to barcodes to build feature-barcode matrices for downstream analysis.

5.4.5.7 scRNA-seq quality control We processed the filtered feature-barcode matrices produced by Cell Ranger in R using `Seurat` [432]. Abnormally low (<200) or high (E12 >9000; E18>7000; P5>7500; P30>4500) numbers of features (Fig. 5.6a-d) indicative of low quality and low information or doublet cells respectively were removed [433, 434]. While a stringent threshold (<5% [435]) for mitochondrial contamination is often recommended, cells with high energy demands may contain higher than expected mitochondrial sequence [436, 437]. Based on the experimental design of capturing actively dividing kidney cells, we loosened this threshold to remove cells with >30% mitochondrial reads (Fig. 5.6e-h) to retain high quality samples. Finally, hemoglobin mapped reads indicative of red blood cell contamination

are restricted to retain cells with <30% hemoglobin (Fig. 5.6i-l).

5.4.5.8 scRNA-seq cell clustering Next, scRNA-seq cells were normalized (see Seurat `NormalizeData`) by log-transforming the scaled (number of features divided by library size and multiplied by a scaling factor of 10000) read counts + 1 [432, 438]. We merged each scRNA-seq time point and identified the top 2000 most variable features (see Seurat `FindVariableFeatures`) between cells. We applied a linear transformation (see Seurat `ScaleData`) to scale expression to a mean of 0 between cells with a maximum variance of 1 to reduce the dominance of highly-expressed genes prior to dimensionality reduction.

Next, we performed Principal Component Analysis (PCA) on the normalized data using the previously identified top features (see Seurat `RunPCA`). To reduce technical noise, we performed a JackStraw [439] procedure (see Seurat `JackStraw`) which randomly subsamples the data, re-performs PCA, and identifies the components with the highest enrichment of low p-value features. We then implemented a graph-based clustering approach to identify groups of similar cells which we subsequently visualized using UMAP (see Seurat `RunUMAP`).

5.4.5.9 scRNA-seq cell identification We annotated individual cell clusters using marker genes defining each cluster. Specifically, the top 25 marker genes for each cluster were identified by finding differentially expressed genes being present in at least 25% of the cells between groups with a 0.25 log fold change between clusters for that gene (`FindAllMarkers`). These uniquely expressed marker genes were used to annotate individual clusters. To perform this annotation, an iterative approach was performed by comparing the top markers to previously identified markers from a series of mouse cell atlases and a commercial database of cell markers (cellKb). First, internally identified markers from internal data was overlapped with top markers from identified clusters. Second, markers were overlapped with cell identity markers from three atlases of developing mouse kidney cells [24, 440, 441]. Third, top identified markers were matched with cell signatures using the commercial CellKb database to create rank ordered lists of top matching gene signatures to cell signatures. The resulting top ranked hits from all three methods were then merged and cell cluster identities manually annotated.

5.4.6 Integrating transciptome and accessibilome

Next, we performed label-transfer to integrate scRNA-seq data with the scATAC-seq data. Anchors between gene scores, a measure of gene expression based on chromatin accessibility, and RNA expression from the scRNA-seq data was combined through the `addGeneIntegrationMatrix` function from ArchR. First, a gene score matrix is generated

by summing reads in peaks with the nearest annotated gene to calculate gene activity by proxy of chromatin accessibility [224]. With this information, cells with the most similar gene score derived expression profile are matched to scRNA-seq expression. This has the intended benefit of labeling ATAC-derived clusters with the RNA-seq based cell annotations. With this integration, we investigated the chromatin accessibility of individual cell types to identify cell-type specific and co-accessible peaks, differentially accessible regions, enriched motifs, and cellular trajectories.

5.4.7 Renin cell differentiation trajectory

Because mature renin cells are so rare, no previous studies have successfully identified markers for this subpopulation. To identify these cells, we performed trajectory inference to uncover the dynamic chromatin and expression changes that lead from Foxd1+ progenitor cells to renin-expressing cells in later development. We calculated a pseudotime trajectory for each cell. Cells early in the renin cell differentiation trajectory were identified by filtering cells on high Foxd1+ expression from RNA-seq data and gene score activity from open chromatin. We then identified cells with high expression (both integrated RNA-seq expression and gene score activity) of the renin marker genes, Ren1 and Akr1b7. Clusters that fulfilled these requirements were utilized as a backbone of ordered vectors and a pseudotime trajectory constructed using ArchR's `addTrajectory` function. This trajectory, and relevant markers, were then visualized on the UMAP embeddings of the LSI and Harmony corrected reduced matrix using the function `plotTrajectory`.

5.4.7.1 Identifying renin cells To determine the identity of the renin-expressing cells, we evaluated markers of mature renin cells and mature smooth muscle cells (SMCs). Both renin cells and mature smooth muscle cells are derived from a shared lineage, and we initially identify clusters of cells with markers for both cell populations. To specifically identify only mature renin cells, we exclude cells with predominant SMC markers. Specifically, we first identify the intersection of cells in the upper quartile of integrated scRNA-seq expression and the upper quartile of gene score activity derived from the scATAC-seq using the renin cell specific markers: Ren1 and Akr1b7. We then identified the upper quartiles from both scRNA-seq expression and gene score activity for the SMC markers: Acta2, Smtn, Tagln, Myh11, and Cnn1. We then took the difference between the identified renin cells and SMCs to create a renin cell exclusive subpopulation, with the remainder representing mature SMCs.

5.4.8 Pairwise comparisons of renin trajectory cells

With our defined trajectory of progenitor cells and fully differentiated renin cells, we performed pairwise comparisons between each cell group along the trajectory. Following the identification of marker peaks, motifs were annotated and differentially enriched motifs identified between clusters. Footprinting analysis was then performed on identified enriched TF motifs. Co-accessible regions and genes were calculated and visualized in the browser with putative motif binding sites. Finally, TFs that were positively correlated between enriched motifs and their matched genes were determined.

5.4.8.1 Renin trajectory marker peaks With the identification of our population of renin cells and the differentiation trajectory defined, we next sought to look which changes in gene expression, regulatory features, and TF motifs are differentially regulated. We subsetted the clusters comprising the renin cell trajectory and first identified the regulatory regions that differentially identify each subpopulation by adding pseudo-bulk replicates that recapitulate the biological variability within each cluster (`addGroupCoverages`) prior to calling peaks (`addReproduciblePeakSet`). To identify features that were differentially expressed or accessible between clusters, the `getMarkerFeatures` (ArchR) function was used and features with a false discovery rate of less than 0.1 and absolute log2 fold changes greater than 1 were identified. We then plotted marker genes based on integrated scRNA-seq and scATAC-seq data on the UMAP embedded reduced dimensionality matrix.

5.4.8.2 Motif annotations and enrichment Next, we determined enriched TF motifs to identify the most active transcription factors in our cell types of interest. We added motif annotations in the enriched peaks using the cisbp [442] database of annotated motifs through the `addMotifAnnotations` function of ArchR. This method utilizes chromVAR [443] to calculate a TF deviation z-score that indicates relative enrichments of motifs in peak regions compared to background [224, 443]. We leverage these deviation scores to uncover differentially accessible regions and motifs between cell clusters of the renin cell trajectory.

5.4.8.3 TF footprinting After identifying enriched motifs, we sought to validate relevant findings by calculating footprints of TFs of interest. To perform footprinting, motif positions were extracted from the renin cell trajectory clusters using `getPositions` (ArchR). These footprints are calculated using the pseudo-bulk replicates of scATAC-seq data along the trajectory and normalized for Tn5 insertion bias. Footprints of enriched motifs in each renin cell trajectory cluster were then plotted using `plotFootprints` (ArchR).

5.4.8.4 Peak co-accessibility and peak to gene links Because peaks with shared accessibility may represent distinct regulatory networks, identifying co-accessible peaks provides a means to uncover features relevant to clusters along the renin cell trajectory. This co-accessibility represents peaks with strong correlation across many cells and can often indicate cell-type specific regulatory regions. We calculate this information with the `addCoAccessibility` function in ArchR. With integrated scRNA-seq data, we additionally look for correlations between both accessibility and gene expression profiles between marker peaks using `getPeak2GeneLinks` at a resolution of 10000 to reduce the total number of links to prevent over-fitting [224]. These peak-gene linkages are more relevant to regulatory relationships because they include not only correlated peaks but genes whose expression is also correlated between cells. The resulting co-accessible peaks and genes are visualized on browser tracks using the `plotBrowserTrack` function. To visualize how accessible regions and gene expression change along the trajectory, we also plot a heatmap of these linked regions and genes using the `plotPeak2GeneHeatmap` function (ArchR).

We extended visualization in browser tracks by also plotting putative TF binding sites in regions of interest by looking for motif occurrence enriched above random using FIMO [444]. TF binding profile position frequency matrices (PFM) for motifs enriched in marker peaks were obtained from the Jaspar database [445]. For a wide region around the Ren1 gene, enriched motif PFMs were used to find individual motif occurrences in this sequence and plotted alongside tracks of chromatin accessibility and co-accessible peaks. These binding sites were loaded into R as individual GRanges objects [446] and concatenated into a single feature object and visualized in the browser by leveraging the `features` parameter of ArchR' `plotBrowserTrack`.

5.4.8.5 Positive transcription factor regulators Since the specific DNA motifs between related TFs are similar, linking individual TF gene expression with motif enrichments enables the identification of positively correlated genes with their matched motif. This is performed by first calculating the most variable TF deviation z-scores between clusters and then correlating those z-scores with the integrated gene expression from the scRNA-seq data (`correlateMatrices`). Positive TF regulators are then those TFs with motif and gene expression correlation greater than 0.5, p-values less than 0.01, and whose maximum difference in z-score between clusters is in the first quartile. We also calculate this with the gene score activity from scATAC-seq to identify positively correlated TFs based on motif enrichment and gene score (`correlateMatrices`). These correlations identify factors playing a central role in renin cell development.

5.5 Supplemental figures

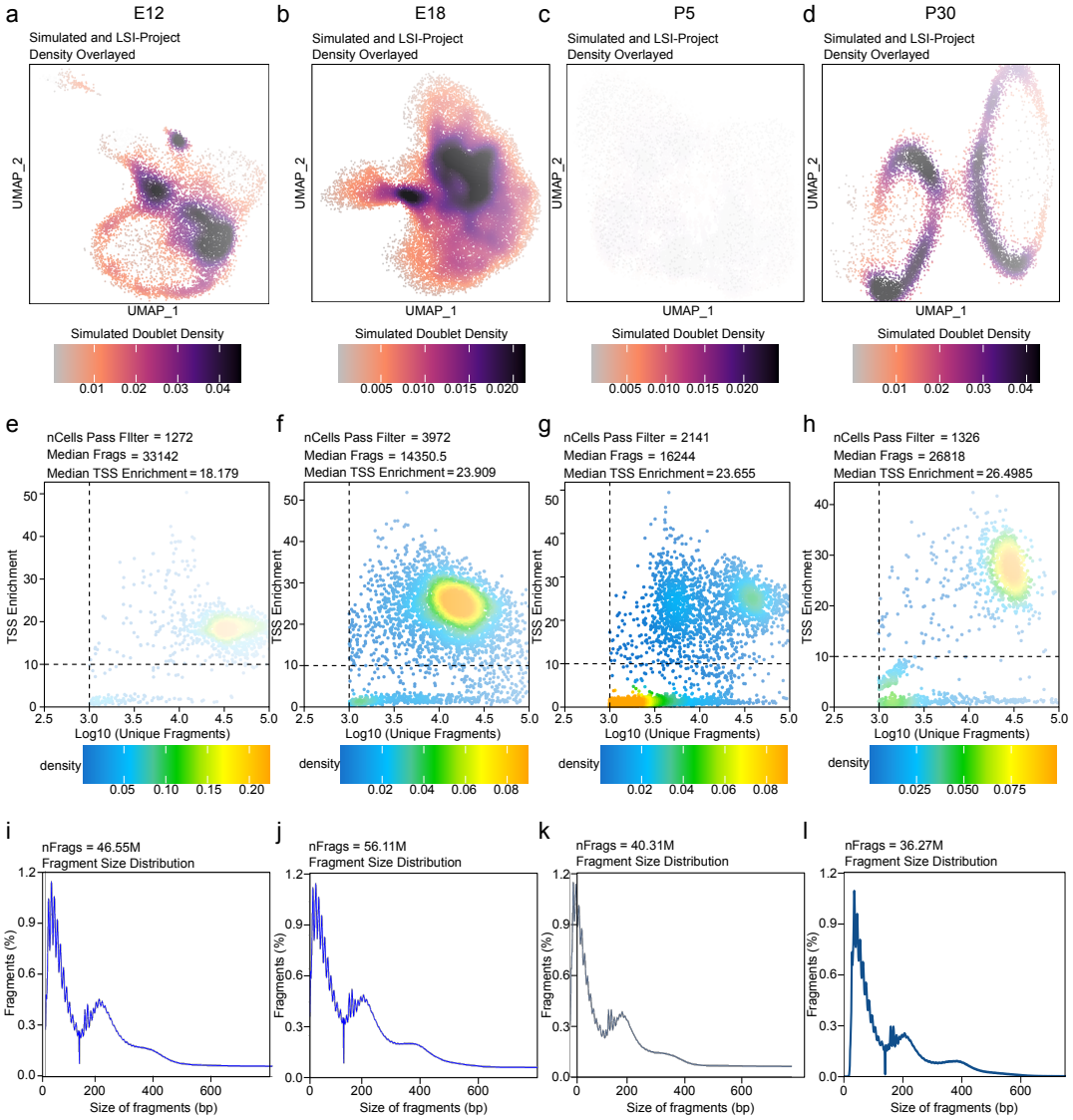

Figure 5.5: **scATAC-seq quality control**. (a-d). UMAP visualiztion of putative doublets in developmental time points E12 (a), E18 (b), P5 (c), and P30 (d). (e-h) Distributions of TSS enrichment by the log10 number of unique fragments for E12 (e), E18 (f), P5 (g), and P30 (h). Dashed lines represent cut off values below which cells are removed. (i-l) Fragment distribution plots for developmental time points E12 (i), E18 (j), P5 (k), and P30 (l).

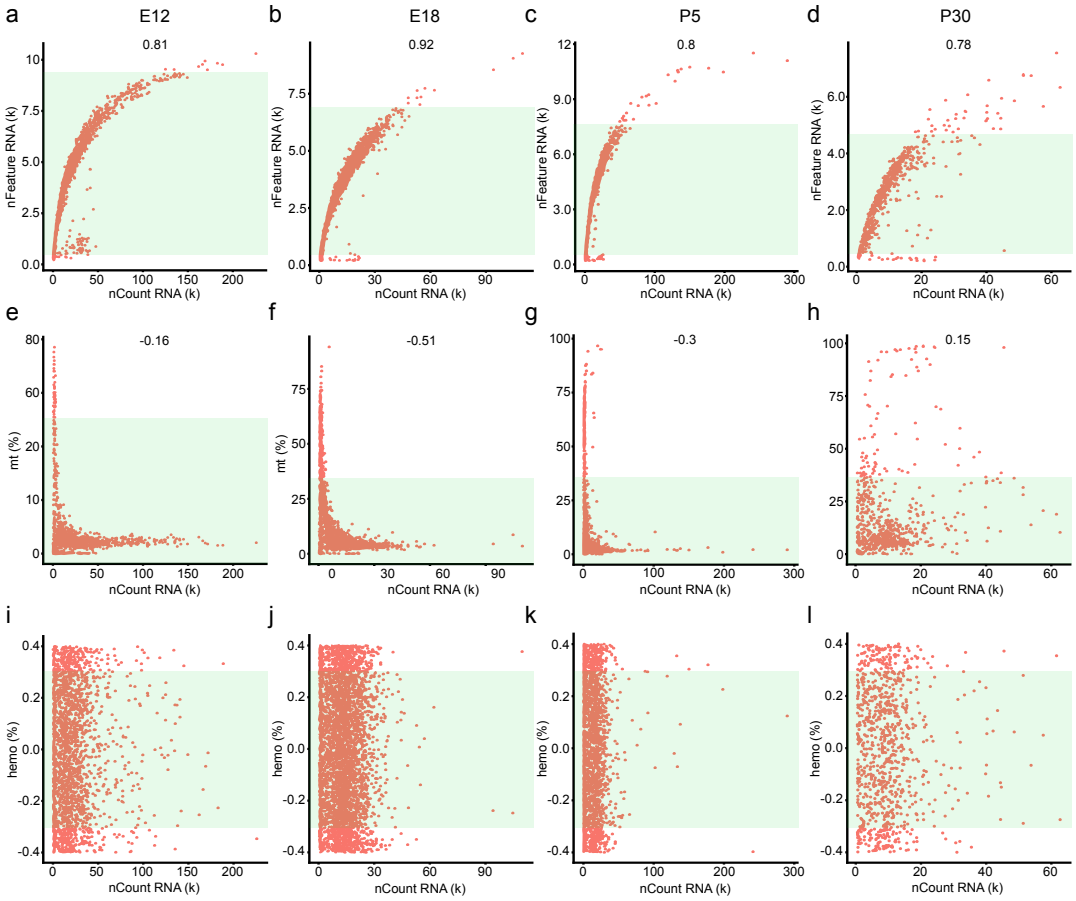

Figure 5.6: **scRNA-seq quality control**. (a-d) Distribution of the number of RNA features against the total RNA count in developmental time points E12 (a), E18 (b), P5 (c), and P30 (d). (e-h) Distribution of the percentage of mitochondria mapped reads against the total RNA count in developmental time points E12 (e), E18 (f), P5 (g), and P30 (h). (i-l) Distribution of the percentage of hemoglobin mapped reads against the total RNA count in developmental time points E12 (i), E18 (j), P5 (k), and P30 (l). Green fill boxes represent cells passing filters.

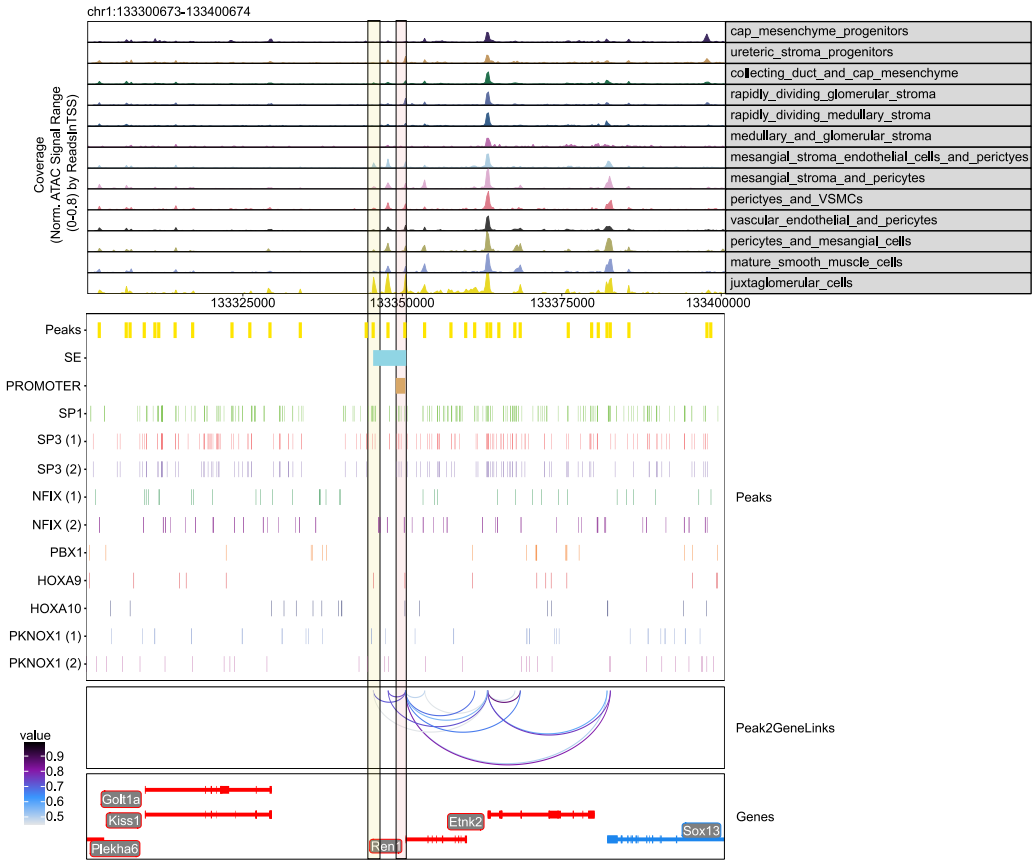

Figure 5.7: **Browser tracks at the Ren1 locus identify previously reported promoter binding factor motif occurrences**. Browser tracks identify preferentially enriched open chromatin in the JG cell cluster at Ren1. Motif occurrences of enriched TFs identify putative binding sites in open and co-accessible peaks. Yellow fill box highlights uniquely enriched peak in JG cluster. Pink fill box highlights JG promoter region.

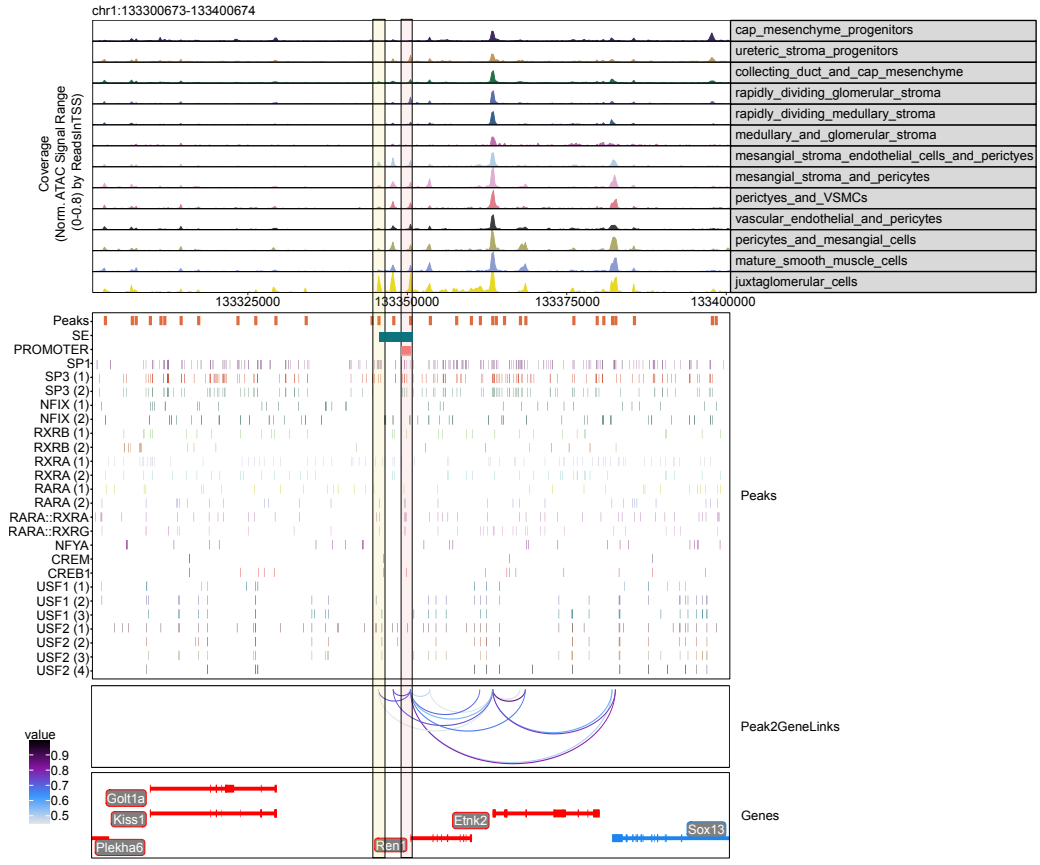

Figure 5.8: **Browser tracks at the Ren1 locus identify previously reported enhancer binding factor motif occurrences**. Browser tracks identify preferentially enriched open chromatin in the JG cell cluster at Ren1. Motif occurrences of enriched TFs identify putative binding sites in open and co-accessible peaks. Yellow fill box highlights uniquely enriched peak in JG cluster. Pink fill box highlights JG promoter region.

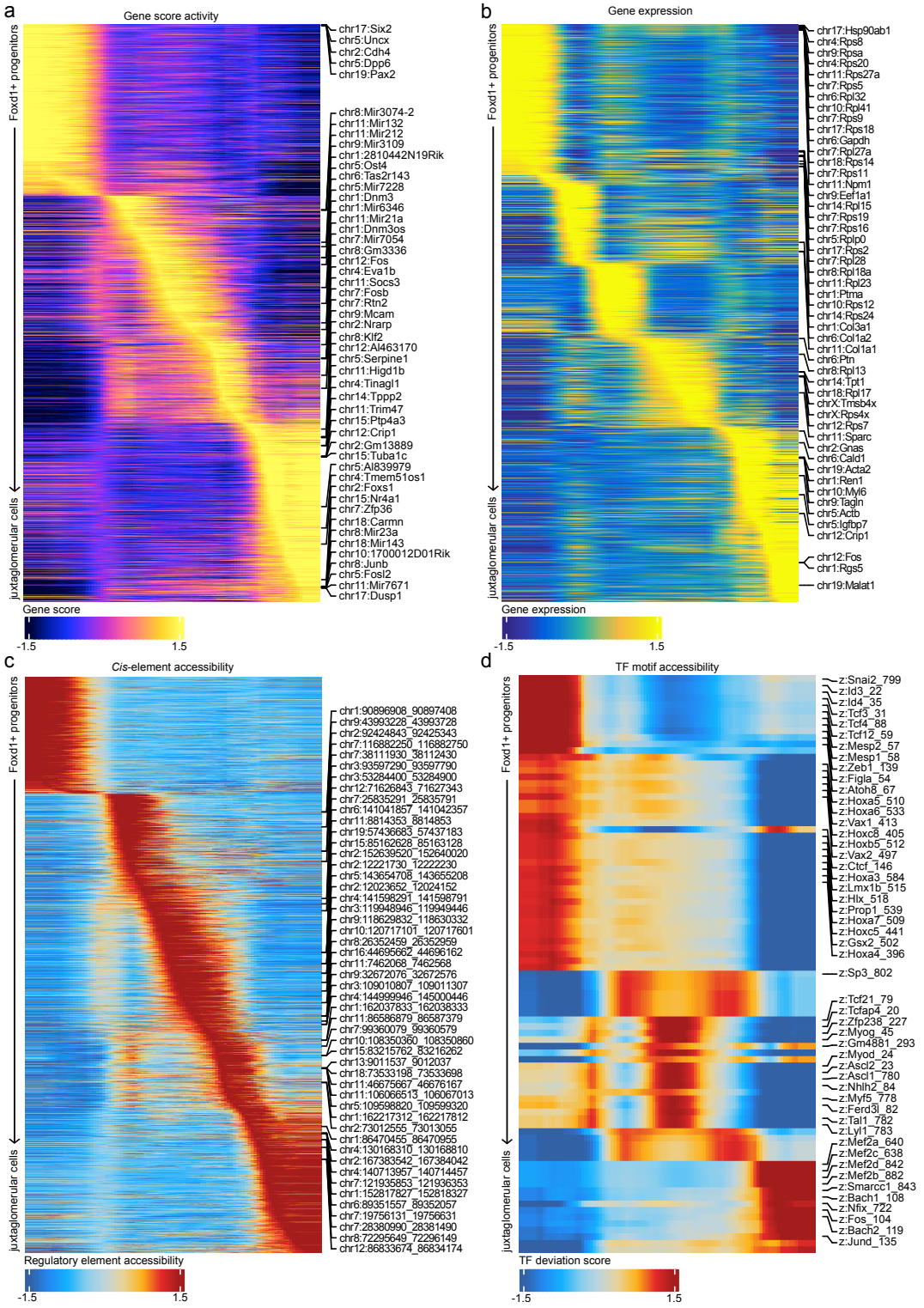

Figure 5.9: **Heatmaps show changes in differential accessibility, expression, and TF motif enrichment along the JG trajectory**. (a) Heatmap of gene score activity along cell clusters defining the JG trajectory. (b) Heatmap of gene expression along cell clusters defining the JG trajectory. (c) Heatmap of accessible regions identified along cell clusters defining the JG trajectory. (d) Heatmap of enriched TF motifs along cell clusters defining the JG trajectory.

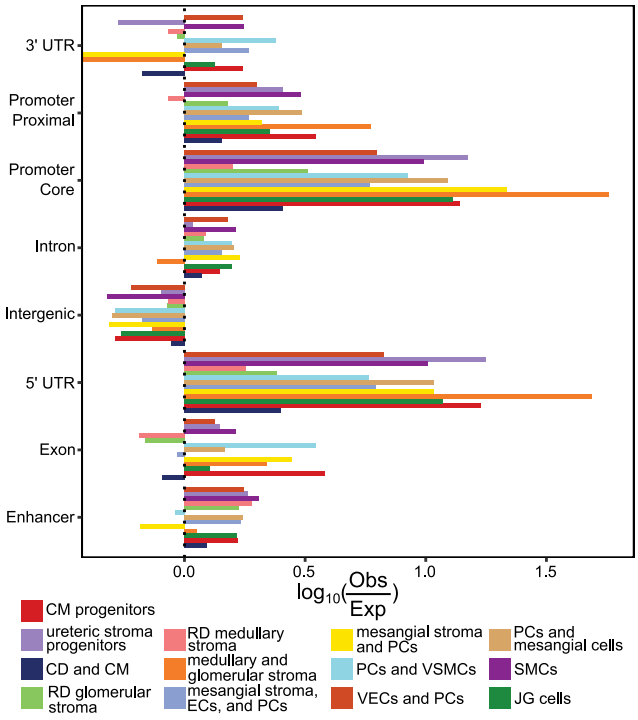

Figure 5.10: **Enrichment of genomic functional classes in marker peaks along the JG differentiation trajectory**. Individual cell clusters along the differentiation trajectory leading to JG cells display differential enrichment of genomic classes including: 3' UTR, promoter proximal, promoter core, intron, intergenic, 5' UTR, exon, and enhancers. Bars for each class represent the observed proportion of regions defined as marker peaks for individual clusters relative to the expected proportion based on the number of bases defined as a particular genomic functional class.

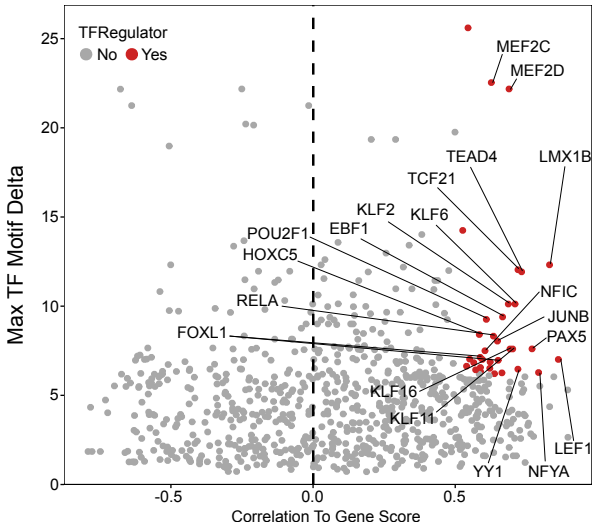

Figure 5.11: **Positive transcription factor regulators of the JG differentiation trajectory** . TFs whose gene expression and motif enrichment are positively correlated indicate putative drivers of differentiation.

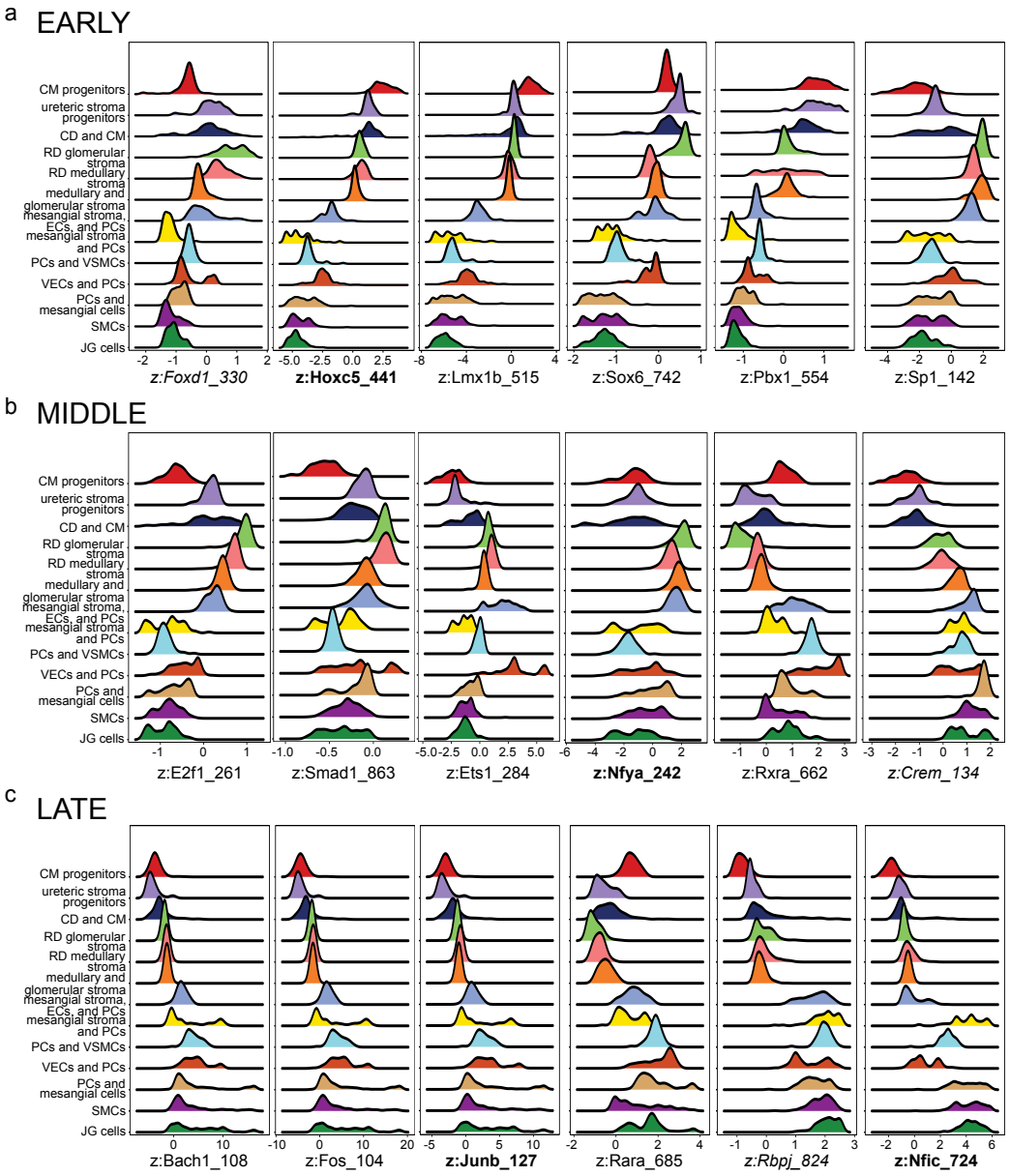

Figure 5.12: **Transcription factors enriched along the JG differentiation trajectory act as early to middle to late acting factors**. TFs that are enriched in clusters along the developmental trajectory of JG cells early (a), middle (b), or late (c). (b) **Bold** TFs are positive TF regulators *Italicized* TFs are known regulators of renin expression. JG: juxtaglomerular

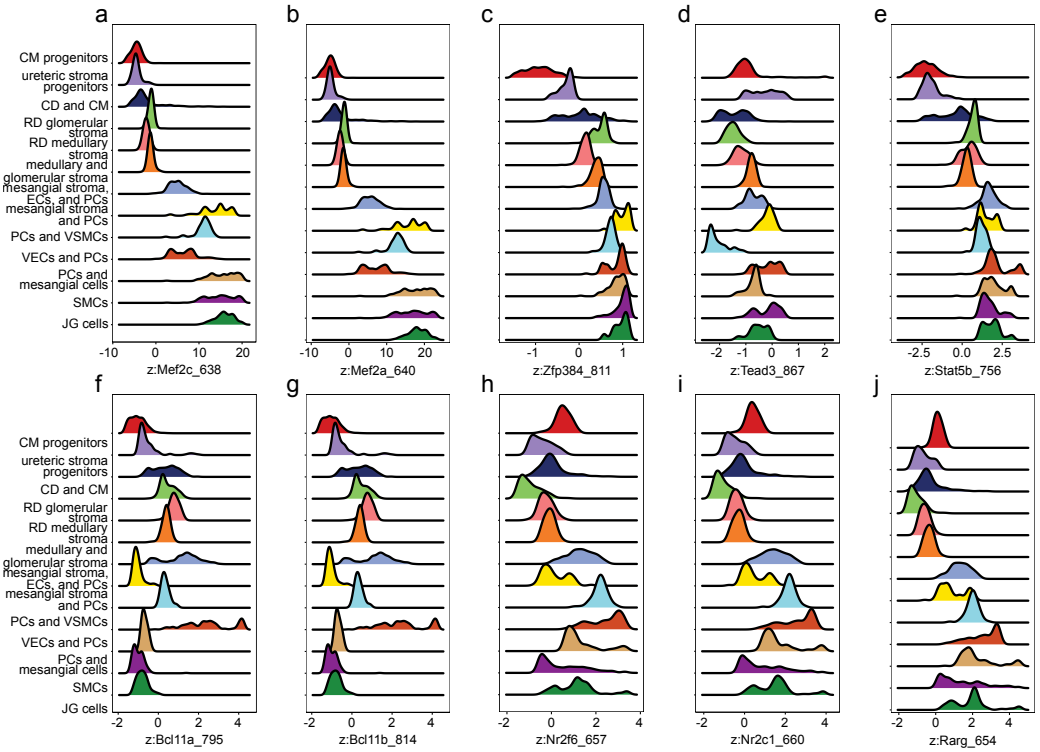

Figure 5.13: **Transcription factors differentiating JG cells from mature SMCs are preferentially enriched in late differentiation trajectory clusters**. (a) Mef2c (and Mef2a (b), Mef2b (not shown), Mef2d (not shown)) are enriched in late time point clusters that contain and ultimately form mature JG cells. Zfp283 (c) is enriched middle to late along the differentiation trajectory. Tead3 (d) differentiates JGs from SMCs but is generally equally enriched across development. Stat5b (e) is positively enriched throughout the middle and late clusters along the JG differentiation trajectory. Bcl11a (f) and Bcl11b (g) differentiate JG cells from SMCs but are overall lowly enriched in all clusters. Nr2f6 (h) and Nr2c1 (i) are enriched in late time point clusters including JG cells. Rarg (j) is enriched in late developmental time points including JG cells. JG: juxtaglomerular

6 Companion research that improves our ability to interrogate genomic regions

While we sought to meet unmet computational needs in the field of genomic region analyses, we realized a significant lack of infrastructure based tools and companion software to streamline and increase reproducibility of bioinformatic analyses existed. To address some of these limitations, I contributed programmatic efforts towards pipeline reporting, genomic asset management, and novel applications of single cell embedding algorithms to improve our ability to uncover biological relationships in region data.

6.1 Looper: A pipeline submission engine

`Looper` is a `Python` package that simplifies job submission to local or cluster-based compute resources. By employing a single definition of project data (*i.e.*, `PEP` formatted files [447]), `looper` handles job submission and status independent of individual pipelines. While the earliest version of `looper` included this utility, it lacked universal reporting features that eased adoption nor organized output in user-friendly and aesthetically pleasing ways. To address this limitation, I wrote updated `Python` functions that could interpret `PEP` projects and pipeline output to generate beautiful HTML-based reports of pipeline output agnostic of individual pipelines. This provides end-users with a browsable report that condenses pipeline output in a single, easy-to-navigate location.

6.2 Refgenie: a reference genome resource manager (derived from [448])

Non de novo sequencing reads require alignment to reference genomes. The challenge arises in which there are multiple reference genome authorities [449–453], but no standardized means to organize and maintain provenance of those assets. `Refgenie` was designed to address this need by providing programmatic access to download and manage assets while using sequence-derived identifiers that uniquely and reproducibly identify asset provenance [448]. Not only does `refgenie` manage primary genome assemblies, but it can build and manage any genome related asset including: gene and transcript annotations, genomic regions annotations, or single-nucleotide polymorphism annotations. To expand the number of managed assets, I wrote `refgenie` recipes to build genomic region annotation assets and genomic indexing assets that identify mappable genomic regions based on read lengths [448].

6.3 Embeddings of genomic region sets capture rich biological associations in lower dimensions (derived from [454])

Genomic regions can represent genes and non-coding regions that include regulatory elements. A common task in genomic analyses involves clustering related genes or regions to find shared patterns of regulation and expression among or between cells and experimental conditions. In natural language processing (NLP) research, significant effort has been performed to design algorithms that can uncover patterns of usage and similarity between words and sentences in language [456]. Perhaps unsurprisingly, these representations of language share strong similarity with representations of the relationships between genomic regions. A word may represent an individual gene or regulatory locus with sentences representing gene and

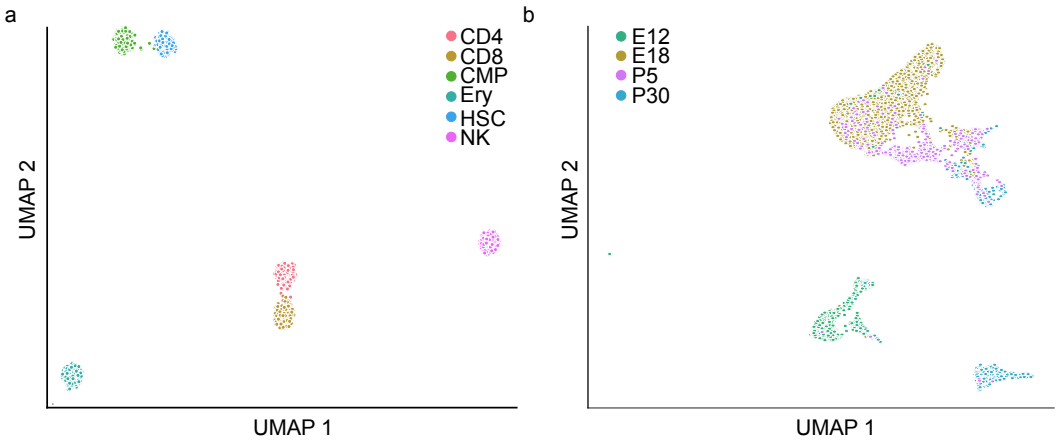

Figure 6.1: **UMAP visualization of scATAC-seq datasets using region-set2vec successfully separates single cells into biologically meaningful clusters.** (a) Simulated bone marrow dataset with a coverage of 2500 fragments per cell [455]. (b) Mouse Foxd1+ progenitors from four developmental time points: E12, E18, P5, and P30.

regulatory element networks. By modifying a well-established NLP algorithm, specifically `Word2Vec` [457], to genomic regions, we showed that meaningful biological relationships can be obtained through this embedding and clustering method [454]. This modified method we term `region-set2vec` to reflect its genomic context. I wrote a `Python` package to apply the `region-set2vec` method to open chromatin and show that NLP-derived algorithms can successfully capture biological variability between groups of scATAC-seq cells with known function (Fig. 6.1a) and/or shared lineage (Fig. 6.1b) [454].

6.4 GenomicDistributions: fast analysis of genomic intervals with bioconductor (derived from [458])

Analyzing genomic regions includes several near-universal procedures which include identifying the distribution of regions across the genome, across annotated functional classes, or relationships between genomic regions and TSSs or genes. With the abundance of open chromatin data rapidly expanding, the need for software capable of processing these large data sets quickly and efficiently continues to be an unmet need. With GenomicDistributions, we sought to address this requirement by producing an R package that is fast and easy to use for the summarization and visualization of genomic regions. Here, I contributed calculation and plotting functions to identify the distribution of genomic regions across annotated functional classes such as promoters, enhancers, exons, introns, or intergenic regions. Furthermore, I included the ability to calculate the expected frequency of these annotation classes over background, emphasing which particular classes are enriched in any provided set of genomic regions. These functions, and others included in the package, enhance our ability to quickly

analyze and visualize any number of large genomic region data sets.

7 Conclusions, Impact, and Future Studies

Chromatin structure and accessibility in cells plays a pivotal role in development, function, and identity. Efforts to investigate chromatin and the epigenome have been rapidly expanding [213] with continual developments in methods to assay regions of open chromatin that include regulatory regions. These regions represent *cis*-regulatory DNA sequences that control the specificity and quantity of transcription through interactions with sequence-specific DNA-binding transcription factors. Chromatin structure also regulates these processes through differences in three-dimensional conformation, nucleosome positioning and density, and through combinations of post-translational modifications of core histone proteins. Combined, open or closed regions of chromatin demonstrate which areas of the genome are transcriptionally active or repressed. Through bench-based and computational analysis, we can study cellular differentiation and function as marked by changes in local chromatin structure.

While historically genome function was interrogated through the linear DNA sequence, we now understand that chromatin accessibility and structure plays pivotal roles in genome regulation. This has significant implications in the study of human health and disease [459–463]. Therefore, we now have the tools to both identify novel druggable targets in diseases, including cancers, and the ability to better understand off-target effects of current therapies. Applying these approaches in understudied systems or in cells historically intractable to traditional methods, such as renin cells, provides the opportunity for novel insights. Here, we have coalesced a number of methods in the field to improve our computational-based tool set and apply these methods to the study of open chromatin in a tissue and cell highly relevant to wide-ranging human health conditions.

7.1 Summary of fulfilled gaps in the field

Each chapter of this dissertation seeks to fulfill or answer an unmet need in the field of bioinformatic analysis. The review of open chromatin analysis provides the field a starting point where new and accomplished users alike can begin when starting a study utilizing ATAC-seq. `PEPPRO` provides the first nascent RNA sequencing pipeline for the field with novel metrics that can be used with `PEPPRO` or independently to validate nascent experiments [215]. `PEPATAC` provides the most comprehensive ATAC-seq pipeline in the field to simplify and universalize ATAC-seq studies in preparation for the plethora of possible downstream investigations [177]. The integration of computational analysis with bench-driven experiments

in the field of renin cell biology provides a detailed example of the power of these assays and tools developed to investigate them. This work demonstrates the utility of multi-omic approaches and how we can apply tools to increase our power of discovery. Finally, the number of companion tools [448, 454] I have helped develop illustrate the necessity for creating a shared infrastructure for discovery and project sharing to aid reproducibility and the robustness of findings. This companion work also uncovers new avenues for genomic regulation discovery [454].

7.2 Integration of computational and bench-based methods

Together, this work brings together the power of computational biology with bench-driven research. It illustrates the necessary steps to train an individual to build software that aids researchers and demonstrates how those tools can then be applied to investigate biological problems that remain challenging to investigate historically. Moving forward, more work remains to improve our analytical tool box, and having researchers trained in both the computational and biological paradigms necessary to answer these questions is an area of work that requires continued effort. This dissertation has revolved exclusively at this intersection: combining biological knowledge and bench-based techniques with the computational power to aid in understanding new findings.

7.3 Support for analytical tools

With the production of publicly-accessible pipelines and tools comes the challenge of long-term maintenance and support. Both `PEPPRO` and `PEPATAC` are published pipelines available through the code sharing platform GitHub [464]. Development has continued regularly post-publication through field-driven changes and community requests. These public code bases to which I can maintain access will enable continued activity, contribution, and support on my part going forward. I will also continue to provide input and support for shared efforts including `looper`, `refgenie`, and the single-cell embeddings tool discussed in Chapter 6. All of the aforementioned tools and analyses are open source and actively participate with users to continue to improve and aid in adoption.

7.4 Investigating the development and regulation of renin cells

Because of the contribution of renin expressing cells in the kidney to homeostasis, improved understanding of the development and regulation of these cells is an ongoing need in the field. While we have continued the effort to explore the genetic and epigenetic landscape of renin expressing cells, future work will expand on this knowledge through several potential avenues

of research.

7.4.1 Epigenetic regulation of early, middle, and late progenitors of renin cells

With a single-cell atlas of chromatin accessibility and transcriptomic changes of renin lineage cells, we now have the ability to investigate differential changes in regulatory element accessibility, gene expression, and TF binding events that are required for differentiation from the cap mesenchyme in developing kidneys to fully mature JG cells in the adult. A likely path forward would include integrating ChIP data and chromosomal conformation experiments to our single-cell developmental atlas. Assaying regulatory histone marks including H3K27ac, H3K27me3, H3K4me1 can enable the identification of when regulatory regions along the trajectory are poised or active to better elucidate the timing of events. This also informs on our ability to obtain a viable cell culture model of renin cells in vitro, where the supplementation of appropriate factors may facilitate long-term culturing and sustainability of renin expression which is currently impossible. We can also integrate chromosome conformation capture with ChIP-seq using HiChiP [465]. Here, we would obtain simultaneous interrogation of where histone marks are identified *and* what DNA loops are associated with those marks of interest.

7.4.2 Identifying druggable targets of aberrant renin cell function and renin expression

With >30% of the population of the United States experiencing hypertension [466, 467], an improved understanding of renin cell function provides valuable insight into the effects of this disease and consequences of standard treatment with RAS inhibitors [391]. Hypertensive individuals constantly stimulate renin production and this has the malignant consequence of leading to renin transformation and arterial disease. Furthermore, inhibition of the angiotensin system to treat hypertension, whether through direct inhibition of renin (aliskiren), by ACE inhibitors, or angiotensin receptor blockers (ARBs) causes significant changes to renin cells [230]. Therefore, our efforts to better understand renin cell function may provide novel targets for hypertensive treatment as well as better understand how current treatments adversely effect renin cells. Future studies will look at possible interactions between drug targets and the identified TFs that contribute to renin cell identity.

8 Materials and Methods

The primary methods for each chapter are included where appropriate. Because this work is computational in nature, it is important to understand the infrastructure utilized and

required to carry out this effort. The logic behind how to tackle computational projects is also discussed.

8.1 Computational infrastructure

This dissertation research is computationally driven and relies on a robust computational infrastructure to be pursued. As a student of Dr. Nathan Sheffield's research group and as a member of the Center for Public Health Genomics (CPHG) at UVA, we have access to a comprehensive computational infrastructure that supported the work of this dissertation. This research utilized multi-display Dell desktop computers and the UVA university-wide computing cluster termed Rivanna (a Cray CS300AC with 240 20-core nodes (4800 cores in total) and two 16-core GPU nodes, all with 128GB of memory per node and a high-speed, low-latency Infiniband interconnect).

The CPHG's linux servers and desktop backups are managed by the UVA Information Technology Services (ITS) group. UVA ITS is an organization of professional faculty and staff dedicated to providing access to high performance computing for research and education. ITS staff manage and maintain the integrated systems infrastructure equipment and networks, UVA email servers, and host a system for the development of dynamic web database applications, designed to support medical research that may need to store sensitive data. These computing resources provide outstanding computational power for bioinformatics. There is extensive support provided for all aspects of computing, including the security of data, backup of data, and 24/7 monitoring of all systems (including dedicated project servers).

For longer running and more intensive parallel-computing analyses, this work relied on access to a university-wide computing cluster, Rivanna, installed in 2014. The system provides 1.4PB of temporary storage in a fast Lustre filesystem. The resource manager is SLURM (Simple Linux Utility for Resource Management). User-level applications are invoked by means of modules. The Cray cluster combines large amounts of processing with large amounts of memory to provide a significant resource for computationally-intensive research at UVA.

8.2 Building a better pipeline

A primary focus of this research has been the development of robust computational pipelines for the analysis of emerging genomic technologies, specifically ATAC-seq and nascent RNA-seq. Much of the bioinformatic software available in the genomic and epigenomic field suffers from a number of common issues, particularly a lack of formal best-practices software training among researchers and limited funding and support for development and maintenance [468–473].

Furthermore, a common question that comes up in any software driven effort is the selection of an appropriate coding environment in which to write such software. To address all of these concerns, firstly, I have performed this work under a computationally trained mentor. Secondly, all implemented code is stored on a publicly-accessible and disclosed platform, GitHub [464], which eases communication from users and developers and streamlines long-term maintenance. Thirdly, I learned and wrote all demonstrated pipelines and tools in either the R or Python coding environment, both of which are the top two most commonly employed bioinformatic languages [474].

9 References

1. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-seq whole-transcriptome analysis of a single cell. Nature methods. 2009;6:377–82.

2. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. Nature biotechnology. 2012;30:777–82.

3. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nature methods. 2013;10:1096–8.

4. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-seq: Single-cell RNA-seq by multiplexed linear amplification. Cell Reports. 2012;2:666–73. doi:https://doi.org/10.1016/j.celrep.2012.08.003.

5. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472:90–4.

6. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science (New York, NY). 2012;338:1622–6.

7. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014;512:155–60.

8. Lohr JG, Adalsteinsson VA, Cibulskis K, Choudhury AD, Rosenberg M, Cruz-Gordillo P, et al. Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. Nature biotechnology. 2014;32:479–84.

9. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, et al. Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. Anal Chem. 2009;81:6813–22. doi:10.1021/ac901049w.

10. Bodenmiller B, Zunder ER, Finck R, Chen TJ, Savig ES, Bruggner RV, et al. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. Nature biotechnology. 2012;30:858–67.

11. Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schönegger A, Klughammer J, et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell reports. 2015;10:1386–97.

12. Guo H, Zhu P, Yan L, Li R, Hu B, Lian Y, et al. The DNA methylation landscape of human early embryos. Nature. 2014;511:606–10. doi:10.1038/nature13544.

13. Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nature biotechnology. 2015;33:1165–72.

14. Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (New York, NY). 2009;326:289–93.

15. Wit E de, Laat W de. A decade of 3C technologies: Insights into nuclear organization. Genes & development. 2012;26:11–24.

16. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, et al. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. Nature. 2013;502:59–64. doi:10.1038/nature12593.

17. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. 2015;348:910–4. doi:10.1126/science.aab1601.

18. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523:486–90.

19. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015;525:251–5.

20. Gaublomme JT, Yosef N, Lee Y, Gertner RS, Yang LV, Wu C, et al. Single-cell genomics unveils critical regulators of Th17 cell pathogenicity. Cell. 2015;163:1400–12.

21. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nature neuroscience. 2016;19:335–46.

22. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nature Biotechnology. 2016;34:1145–60. doi:10.1038/nbt.3711.

23. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566:496–502. doi:10.1038/s41586-019-0969-x.

24. Combes AN, Phipson B, Lawlor KT, Dorison A, Patrick R, Zappia L, et al. Single cell analysis of the developing mouse kidney provides deeper insight into marker gene expression and ligand-receptor crosstalk. Development (Cambridge, England). 2019;146.

25. Cairns BR. The logic of chromatin architecture and remodelling at promoters. Nature. 2009;461:193–8.

26. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nature Reviews Genetics. 2019;20:207–20. doi:10.1038/s41576-018-0089-8.

27. Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. Nature reviews Genetics. 2016;17:487–500.

28. Dann GP, Liszczak GP, Bagert JD, Müller MM, Nguyen UTT, Wojcik F, et al. ISWI chromatin remodellers sense nucleosome modifications to determine substrate preference.

Nature. 2017;548:607–11.

29. Göndör A, Ohlsson R. Chromosome crosstalk in three dimensions. Nature. 2009;461:212–7.

30. Cavalli G, Misteli T. Functional implications of genome topology. Nature structural & molecular biology. 2013;20:290–9.

31. Gibcus JH, Dekker J. The hierarchy of the 3D genome. Molecular cell. 2013;49:773–82.

32. Cao J, Luo Z, Cheng Q, Xu Q, Zhang Y, Wang F, et al. Three-dimensional regulation of transcription. Protein & cell. 2015;6:241–53.

33. Woodcock CL. Chromatin architecture. Current opinion in structural biology. 2006;16:213–20.

34. Li B, Carey M, Workman JL. The role of chromatin during transcription. Cell. 2007;128:707–19.

35. Ahmad K, Henikoff S. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. Molecular cell. 2002;9:1191–200.

36. Huisinga KL, Brower-Toland B, Elgin SCR. The contradictory definitions of heterochromatin: Transcription and silencing. Chromosoma. 2006;115:110–22.

37. Grewal SIS, Jia S. Heterochromatin revisited. Nature reviews Genetics. 2007;8:35–46.

38. Allshire RC, Madhani HD. Ten principles of heterochromatin formation and function. Nature reviews Molecular cell biology. 2018;19:229–44.

39. Martire S, Banaszynski LA. The roles of histone variants in fine-tuning chromatin organization and function. Nature reviews Molecular cell biology. 2020;21:522–41.

40. Levine M, Tjian R. Transcription regulation and animal diversity. Nature. 2003;424:147–51.

41. Hager GL, McNally JG, Misteli T. Transcription dynamics. Molecular cell. 2009;35:741–53.

42. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009;461:199–205.

43. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, et al. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012;488:116–20.

44. Dean A. In the loop: Long range chromatin interactions and gene regulation. Briefings in functional genomics. 2011;10:3–10.

45. Laat W de, Duboule D. Topology of mammalian developmental enhancers and their regulatory landscapes. Nature. 2013;502:499–506.

46. Kornberg RD. Chromatin structure: A repeating unit of histones and DNA. Science (New York, NY). 1974;184:868–71.

47. Kornberg RD, Lorch Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell. 1999;98:285–94.

48. Davey CA, Sargent DF, Luger K, Maeder AW, Richmond TJ. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. Journal of molecular biology. 2002;319:1097–113.

49. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature. 2009;458:362–6.

50. Lee C-K, Shibata Y, Rao B, Strahl BD, Lieb JD. Evidence for nucleosome depletion at active regulatory regions genome-wide. Nature genetics. 2004;36:900–5.

51. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82. doi:10.1038/nature11232.

52. Knezetic JA, Luse DS. The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. Cell. 1986;45:95–104.

53. Lorch Y, LaPointe JW, Kornberg RD. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. Cell. 1987;49:203–10.

54. Khorasanizadeh S. The nucleosome: From genomic organization to genomic regulation. Cell. 2004;116:259–72.

55. Radman-Livaja M, Rando OJ. Nucleosome positioning: How is it established, and why does it matter? Developmental biology. 2010;339:258–66.

56. Woodcock CL, Safer JP, Stanchfield JE. Structural repeating units in chromatin. I. Evidence for their general occurrence. Experimental cell research. 1976;97:101–10.

57. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 a resolution. Nature. 1997;389:251–60.

58. Johnson SM, Tan FJ, McCullough HL, Riordan DP, Fire AZ. Flexibility and constraint in the nucleosome core landscape of caenorhabditis elegans chromatin. Genome research. 2006;16:1505–16.

59. Mariño-Ramírez L, Kann MG, Shoemaker BA, Landsman D. Histone structure and nucleosome stability. Expert review of proteomics. 2005;2:719–29.

60. Kimura H. Histone modifications for human epigenome analysis. Journal of human genetics. 2013;58:439–45.

61. Harr JC, Gonzalez-Sandoval A, Gasser SM. Histones and histone modifications in perinuclear chromatin anchoring: From yeast to man. EMBO reports. 2016;17:139–55.

62. Venkatesh S, Workman JL. Histone exchange, chromatin structure and the regulation of transcription. Nature reviews Molecular cell biology. 2015;16:178–89.

63. Creyghton MP, Markoulaki S, Levine SS, Hanna J, Lodato MA, Sha K, et al. H2AZ is enriched at polycomb complex target genes in ES cells and is necessary for lineage commitment.

Cell. 2008;135:649–61.

64. Hu G, Cui K, Northrup D, Liu C, Wang C, Tang Q, et al. H2A.z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. Cell stem cell. 2013;12:180–92.

65. Liang G, Lin JCY, Wei V, Yoo C, Cheng JC, Nguyen CT, et al. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. Proceedings of the National Academy of Sciences of the United States of America. 2004;101:7357–62.

66. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129:823–37.

67. Calo E, Wysocka J. Modification of enhancer chromatin: What, how, and why? Molecular cell. 2013;49:825–37.

68. Dong X, Weng Z. The correlation between histone modifications and gene expression. In: Epigenomics. 2013. p. 113–6.

69. Nguyen MLT, Jones SA, Prier JE, Russ BE. Transcriptional enhancers in the regulation of t cell differentiation. Frontiers in immunology. 2015;6:462.

70. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nature Reviews Molecular Cell Biology. 2018;19:621–37. doi:10.1038/s41580-018-0028-8.

71. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature genetics. 2007;39:311–8.

72. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences. 2010;107:21931–6. doi:10.1073/pnas.1016071107.

73. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9.

74. Crispatzu G, Rehimi R, Pachano T, Bleckwehl T, Cruz-Molina S, Xiao C, et al. The chromatin, topological and regulatory properties of pluripotency-associated poised enhancers are conserved in vivo. Nature Communications. 2021;12:4344. doi:10.1038/s41467-021-24641-4.

75. Natoli G, Ghisletti S, Barozzi I. The genomic landscapes of inflammation. Genes & development. 2011;25:101–6.

76. Cruz-Molina S, Respuela P, Tebartz C, Kolovos P, Nikolic M, Fueyo R, et al. PRC2 facilitates the regulatory topology required for poised enhancer function during pluripotent

stem cell differentiation. Cell stem cell. 2017;20:689–705.e9.

77. Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature. 2013;499:360–3.

78. Ntini E, Järvelin AI, Bornholdt J, Chen Y, Boyd M, Jørgensen M, et al. Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nature structural & molecular biology. 2013;20:923–8.

79. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature genetics. 2014;46:1311–20.

80. Riethoven J-JM. Regulatory regions in DNA: Promoters, enhancers, silencers, and insulators. In: Ladunga I, editor. Computational biology of transcription factor binding. Totowa, NJ: Humana Press; 2010. p. 33–42. doi:10.1007/978-1-60761-854-6_3.

81. Chatterjee S, Ahituv N. Gene regulatory elements, major drivers of human disease. Annual review of genomics and human genetics. 2017;18:45–63.

82. Dynlacht BD. Regulation of transcription by proteins that control the cell cycle. Nature. 1997;389:149–52.

83. Ptashne M, Gann A. Transcriptional activation by recruitment. Nature. 1997;386:569–77.

84. Takagi Y, Kornberg RD. Mediator as a general transcription factor. The Journal of biological chemistry. 2006;281:80–9.

85. Kadonaga JT. Perspectives on the RNA polymerase II core promoter. Wiley interdisciplinary reviews Developmental biology. 2012;1:40–51.

86. Spitz F, Furlong EEM. Transcription factors: From enhancer binding to developmental control. Nature reviews Genetics. 2012;13:613–26.

87. Ponjavic J, Lenhard B, Kai C, Kawai J, Carninci P, Hayashizaki Y, et al. Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. Genome biology. 2006;7:R78.

88. Smale ST, Baltimore D. The "initiator" as a transcription control element. Cell. 1989;57:103–13.

89. Burke TW, Kadonaga JT. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. Genes & development. 1996;10:711–24.

90. Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH. New core promoter element in RNA polymerase II-dependent transcription: Sequence-specific DNA binding by transcription factor IIB. Genes & development. 1998;12:34–44.

91. Deng W, Roberts SGE. A core promoter element downstream of the TATA box that is

recognized by TFIIB. Genes & development. 2005;19:2418–23.

92. Lewis BA, Kim TK, Orkin SH. A downstream element in the human beta-globin promoter: Evidence of extended sequence-specific transcription factor IID contacts. Proceedings of the National Academy of Sciences of the United States of America. 2000;97:7172–7.

93. Rach EA, Winter DR, Benjamin AM, Corcoran DL, Ni T, Zhu J, et al. Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. PLoS genetics. 2011;7:e1001274.

94. Roider HG, Lenhard B, Kanhere A, Haas SA, Vingron M. CpG-depleted promoters harbor tissue-specific transcription factor binding signals–implications for motif overrepresentation analyses. Nucleic acids research. 2009;37:6305–15.

95. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: Emerging characteristics and insights into transcriptional regulation. Nature reviews Genetics. 2012;13:233–45.

96. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nature genetics. 2006;38:626–35.

97. Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, et al. Genome-wide analysis of promoter architecture in drosophila melanogaster. Genome research. 2011;21:182–92.

98. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006;125:315–26.

99. Akalin A, Fredman D, Arner E, Dong X, Bryne JC, Suzuki H, et al. Transcriptional features of genomic regulatory blocks. Genome biology. 2009;10:R38.

100. Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell. 1981;27:299–308.

101. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: From properties to genome-wide predictions. Nature reviews Genetics. 2014;15:272–86.

102. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nature genetics. 2008;40:897–903.

103. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America. 2010;107:21931–6.

104. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. A unique chro-

matin signature uncovers early developmental enhancers in humans. Nature. 2011;470:279–83.

105. Zentner GE, Tesar PJ, Scacheri PC. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome research. 2011;21:1273–83.

106. Ghavi-Helm Y, Klein FA, Pakozdi T, Ciglar L, Noordermeer D, Huber W, et al. Enhancer loops appear stable during development and are associated with paused polymerase. Nature. 2014;512:96–100. doi:10.1038/nature13417.

107. Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, et al. Functional and topological characteristics of mammalian regulatory domains. Genome research. 2014;24:390–400.

108. Merkenschlager M, Nora EP. CTCF and cohesin in genome folding and transcriptional gene regulation. Annual review of genomics and human genetics. 2016;17:17–43.

109. Spitz F. Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. Seminars in cell & developmental biology. 2016;57:57–67.

110. Zabidi MA, Stark A. Regulatory enhancer-core-promoter communication via transcription factors and cofactors. Trends in genetics : TIG. 2016;32:801–14.

111. Ong C-T, Corces VG. Enhancer function: New insights into the regulation of tissue-specific gene expression. Nature reviews Genetics. 2011;12:283–93.

112. Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. Cell. 2011;144:327–39.

113. Thanos D, Maniatis T. Virus induction of human IFNB gene expression requires the assembly of an enhanceosome. Cell. 1995;83:1091–100. doi:https://doi.org/10.1016/0092-8674(95)90136-1.

114. Reményi A, Schöler HR, Wilmanns M. Combinatorial control of gene expression. Nature Structural & Molecular Biology. 2004;11:812–5. doi:10.1038/nsmb820.

115. Osterwalder M, Barozzi I, Tissières V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, et al. Enhancer redundancy provides phenotypic robustness in mammalian development. Nature. 2018;554:239–43.

116. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153:307–19.

117. Pott S, Lieb JD. What are super-enhancers? Nature Genetics. 2015;47:8–12. doi:10.1038/ng.3167.

118. Phillips JE, Corces VG. CTCF: Master weaver of the genome. Cell. 2009;137:1194–211. doi:https://doi.org/10.1016/j.cell.2009.06.001.

119. Merkenschlager M, Odom DT. CTCF and cohesin: Linking gene regulatory elements

with their targets. Cell. 2013;152:1285–97.

120. Ong C-T, Corces VG. CTCF: An architectural protein bridging genome topology and function. Nature Reviews Genetics. 2014;15:234–46. doi:10.1038/nrg3663.

121. Bell AC, West AG, Felsenfeld G. Insulators and boundaries: Versatile regulatory elements in the eukaryotic genome. Science (New York, NY). 2001;291:447–50.

122. Bushey AM, Dorman ER, Corces VG. Chromatin insulators: Regulatory mechanisms and epigenetic inheritance. Molecular cell. 2008;32:1–9.

123. Riethoven J-JM. Regulatory regions in DNA: Promoters, enhancers, silencers, and insulators. Methods in molecular biology (Clifton, NJ). 2010;674:33–42.

124. Lanzuolo C, Roure V, Dekker J, Bantignies F, Orlando V. Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. Nature Cell Biology. 2007;9:1167–74. doi:10.1038/ncb1637.

125. Tiwari VK, McGarvey KM, Licchesi JDF, Ohm JE, Herman JG, Schübeler D, et al. PcG proteins, DNA methylation, and gene repression by chromatin looping. PLoS biology. 2008;6:2911–27.

126. Harris MB, Mostecki J, Rothman PB. Repression of an interleukin-4-responsive promoter requires cooperative BCL-6 function*. Journal of Biological Chemistry. 2005;280:13114–21. doi:https://doi.org/10.1074/jbc.M412649200.

127. Li L, He S, Sun J-M, Davie JR. Gene regulation by Sp1 and Sp3. Biochemistry and cell biology = Biochimie et biologie cellulaire. 2004;82:460–71.

128. Srinivasan L, Atchison ML. YY1 DNA binding and PcG recruitment requires CtBP. Genes & development. 2004;18:2596–601.

129. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. Annual Review of Genomics and Human Genetics. 2006;7:29–59. doi:10.1146/annurev.genom.7.080505.115623.

130. Darnell JEJ. Variety in the level of gene control in eukaryotic cells. Nature. 1982;297:365–71.

131. Ptashne M. How eukaryotic transcriptional activators work. Nature. 1988;335:683–9.

132. Latchman DS. Transcription factors: An overview. The international journal of biochemistry & cell biology. 1997;29:1305–12.

133. Karin M. Too many transcription factors: Positive and negative interactions. The New biologist. 1990;2:126–31.

134. Wärnmark A, Treuter E, Wright APH, Gustafsson J-A. Activation functions 1 and 2 of nuclear receptors: Molecular strategies for transcriptional activation. Molecular endocrinology (Baltimore, Md). 2003;17:1901–9.

135. Workman JL, Kingston RE. Nucleosome core displacement in vitro via a metastable transcription factor-nucleosome complex. Science (New York, NY). 1992;258:1780–4.

136. Svaren J, Klebanow E, Sealy L, Chalkley R. Analysis of the competition between nucleosome formation and transcription factor binding. The Journal of biological chemistry. 1994;269:9335–44.

137. Krebs AR, Imanci D, Hoerner L, Gaidatzis D, Burger L, Schübeler D. Genome-wide single-molecule footprinting reveals high RNA polymerase II turnover at paused promoters. Molecular cell. 2017;67:411–422.e4.

138. Almer A, Rudolph H, Hinnen A, Hörz W. Removal of positioned nucleosomes from the yeast PHO5 promoter upon PHO5 induction releases additional upstream activating DNA elements. The EMBO journal. 1986;5:2689–96.

139. Taberlay PC, Kelly TK, Liu C-C, You JS, De Carvalho DD, Miranda TB, et al. Polycomb-repressed genes have permissive enhancers that initiate reprogramming. Cell. 2011;147:1283–94.

140. Bao X, Rubin AJ, Qu K, Zhang J, Giresi PG, Chang HY, et al. A novel ATAC-seq approach reveals lineage-specific reinforcement of the open chromatin landscape via cooperation between BAF and p63. Genome biology. 2015;16:284.

141. Swinstead EE, Paakinaho V, Presman DM, Hager GL. Pioneer factors and ATP-dependent chromatin remodeling factors interact dynamically: A new perspective: Multiple transcription factors can effect chromatin pioneer functions through dynamic interactions with ATP-dependent chromatin remodeling factors. BioEssays : news and reviews in molecular, cellular and developmental biology. 2016;38:1150–7.

142. Taylor IC, Workman JL, Schuetz TJ, Kingston RE. Facilitated binding of GAL4 and heat shock factor to nucleosomal templates: Differential function of DNA-binding domains. Genes & development. 1991;5:1285–98.

143. McPherson CE, Shim EY, Friedman DS, Zaret KS. An active tissue-specific enhancer and bound transcription factors existing in a precisely positioned nucleosomal array. Cell. 1993;75:387–98.

144. Steger DJ, Workman JL. Stable co-occupancy of transcription factors and histones at the HIV-1 enhancer. The EMBO journal. 1997;16:2463–72.

145. Cirillo LA, Lin FR, Cuesta I, Friedman D, Jarnik M, Zaret KS. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. Molecular cell. 2002;9:279–89.

146. Zaret KS, Carroll JS. Pioneer transcription factors: Establishing competence for gene expression. Genes & development. 2011;25:2227–41.

147. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, Hoff JP van, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nature biotechnology. 2014;32:171–8.

148. Soufi A, Garcia MF, Jaroszewicz A, Osman N, Pellegrini M, Zaret KS. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell. 2015;161:555–68.

149. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee B-K, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome research. 2011;21:1757–67. doi:10.1101/gr.121541.111.

150. Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, et al. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome research. 2013;23:777–88. doi:10.1101/gr.152140.112.

151. Pálfy M, Schulze G, Valen E, Vastenhouw NL. Chromatin accessibility established by Pou5f3, Sox19b and nanog primes genes for activity during zebrafish genome activation. PLOS Genetics. 2020;16:e1008546. doi:10.1371/journal.pgen.1008546.

152. Tewari AK, Yardimci GG, Shibata Y, Sheffield NC, Song L, Taylor BS, et al. Chromatin accessibility reveals insights into androgen receptor activation and transcriptional specificity. Genome Biol. 2012;13:R88. doi:10.1186/gb-2012-13-10-r88.

153. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence M, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015;518:360–4. doi:10.1038/nature14221.

154. Spivakov M, Fraser P. Defining cell type with chromatin profiling. Nature biotechnology. 2016;34:1126–8. doi:10.1038/nbt.3724.

155. Wang J, Zibetti C, Shang P, Sripathi SR, Zhang P, Cano M, et al. ATAC-seq analysis reveals a widespread decrease of chromatin accessibility in age-related macular degeneration. 2018;9. doi:10.1038/s41467-018-03856-y.

156. Hatzi K, Geng H, Doane AS, Meydan C, LaRiviere R, Cardenas M, et al. Histone demethylase LSD1 is required for germinal center formation and BCL6-driven lymphomagenesis. 2019;20:86–96. doi:10.1038/s41590-018-0273-1.

157. Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, Zaretsky I, Jaitin DA, David E, et al. Chromatin state dynamics during blood formation. 2014;345:943–9. doi:10.1126/science.1256271.

158. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nature genetics. 2016;48:1193–203. doi:10.1038/ng.3646.

159. Chan HL, Beckedorff F, Zhang Y, Garcia-Huidobro J, Jiang H, Colaprico A, et al. Polycomb complexes associate with enhancers and promote oncogenic transcriptional programs in cancer through multiple mechanisms. 2018;9. doi:10.1038/s41467-018-05728-x.

160. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science (New York, NY). 2018;362. doi:10.1126/science.aav1898.

161. Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, Wang Z, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell. 2008;132:887–98.

162. Cui K, Zhao K. Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-seq. Methods in molecular biology (Clifton, NJ). 2012;833:413–9.

163. Klein DC, Hainer SJ. Genomic methods in profiling DNA accessibility and factor localization. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology. 2020;28:69–85.

164. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome research. 2006;16:123–31.

165. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132:311–22.

166. Song L, Crawford GE. DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harbor protocols. 2010;2010:pdb.prot5384.

167. Kumar V, Muratani M, Rayan NA, Kraus P, Lufkin T, Ng HH, et al. Uniform, optimal signal processing of mapped deep-sequencing data. Nature biotechnology. 2013;31:615–22.

168. Simon JM, Giresi PG, Davis IJ, Lieb JD. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. Nature protocols. 2012;7:256–67.

169. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. Genome research. 2007;17:877–85.

170. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. Nature methods. 2013;10:1213. doi:10.1038/nmeth.2688.

171. Tsompana M, Buck MJ. Chromatin accessibility: A window into the genome. Epigenetics & chromatin. 2014;7:33.

172. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee B-K, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity.

Genome research. 2011;21:1757–67.

173. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A method for assaying chromatin accessibility genome-wide. Current protocols in molecular biology. 2015;109:21.29.1–9. doi:10.1002/0471142727.mb2129s109.

174. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Scientific reports. 2017;14:959–62. doi:10.1038/nmeth.4396.

175. Reznikoff WS. Transposon Tn5. Annu Rev Genet. 2008;42:269–86. doi:10.1146/annurev.genet.42.110807.09

176. Adey A, Morrison HG, name) A (no last, Xun X, Kitzman JO, Turner EH, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. 2010;11:R119. doi:10.1186/gb-2010-11-12-r119.

177. Smith JP, Corces MR, Xu J, Reuter VP, Chang HY, Sheffield NC. PEPATAC: an optimized pipeline for ATAC-seq data analysis with serial alignments. NAR Genomics and Bioinformatics. 2021;3. doi:10.1093/nargab/lqab101.

178. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. 2015;523:486–90. doi:10.1038/nature14590.

179. Lareau CA, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nature Biotechnology. 2019;37:916–24.

180. Schwartzman O, Tanay A. Single-cell epigenomics: Techniques and emerging applications. Nature Reviews Genetics. 2015;16:716. doi:10.1038/nrg3980.

181. Fang R, Preissl S, Li Y, Hou X, Lucero J, Wang X, et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nature Communications. 2021;12:1337. doi:10.1038/s41467-021-21583-9.

182. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science (New York, NY). 2008;320:1344–9.

183. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008;453:1239–43.

184. Wang Z, Gerstein M, Snyder M. RNA-seq: A revolutionary tool for transcriptomics. Nature reviews Genetics. 2009;10:57–63.

185. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA

fragments. Nature protocols. 2012;7:1534–50.

186. Baker M. MicroRNA profiling: Separating signal from noise. Nature methods. 2010;7:687–92.

187. Behrens A, Rodschinka G, Nedialkova DD. High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. Molecular cell. 2021;81:1802–1815.e7.

188. Oh S, Song S, Grabowski G, Zhao H, Noonan JP. Time series expression analyses using RNA-seq: A statistical approach. BioMed research international. 2013;2013:203681.

189. Äijö T, Butty V, Chen Z, Salo V, Tripathi S, Burge CB, et al. Methods for time series analysis of RNA-seq data with application to human Th17 cell differentiation. Bioinformatics (Oxford, England). 2014;30:i113–20.

190. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. Briefings in Functional Genomics. 2014;14:130–42. doi:10.1093/bfgp/elu035.

191. Heyer EE, Deveson IW, Wooi D, Selinger CI, Lyons RJ, Hayes VM, et al. Diagnosis of fusion genes using targeted RNA sequencing. Nature Communications. 2019;10:1388. doi:10.1038/s41467-019-09374-9.

192. Halperin RF, Hegde A, Lang JD, Raupach EA, Narayanan V, Huentelman M, et al. Improved methods for RNAseq-based alternative splicing analysis. Scientific Reports. 2021;11:10740. doi:10.1038/s41598-021-89938-2.

193. Maamar H, Raj A, Dubnau D. Noise in gene expression determines cell fate in bacillus subtilis. Science (New York, NY). 2007;317:526–9.

194. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatic-spipelines. Experimental & Molecular Medicine. 2018;50:1–4. doi:10.1038/s12276-018-0071-8.

195. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nature methods. 2014;11:163–6.

196. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14.

197. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161:1187–201.

198. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology. 2014;32:381–6.

199. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, et al.

Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature. 2014;509:371–5.

200. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nature biotechnology. 2015;33:155–60.

201. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nature reviews Genetics. 2015;16:133–45.

202. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008;322:1845–8.

203. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. 2013;339:950–3. doi:10.1126/science.1229386.

204. Chu T, Rice EJ, Booth GT, Salamanca HH, Wang Z, Core LJ, et al. Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. Nature genetics. 2018;1.

205. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465:182–7. doi:10.1038/nature09033.

206. De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, et al. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS biology. 2010;8:e1000384.

207. Koch F, Fenouil R, Gut M, Cauchy P, Albert TK, Zacarias-Cabeza J, et al. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. Nature structural & molecular biology. 2011;18:956–63.

208. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61.

209. Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science (New York, NY). 2015;347:1010–4.

210. Scruggs BS, Gilchrist DA, Nechaev S, Muse GW, Burkholder A, Fargo DC, et al. Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. 2015;58:1101–12. doi:10.1016/j.molcel.2015.04.006.

211. Wang J, Zhao Y, Zhou X, Hiebert SW, Liu Q, Shyr Y. Nascent RNA sequencing analysis provides insights into enhancer-mediated gene regulation. BMC Genomics. 2018;19:633. doi:10.1186/s12864-018-5016-z.

212. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biology. 2016;17.

doi:10.1186/s13059-016-0881-8.

213. Smith JP, Sheffield NC. Analytical approaches for ATAC-seq data analysis. Current Protocols in Human Genetics. 2020;106:e101. doi:10.1002/cphg.101.

214. Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: A hitchhiker's guide to ATAC-seq data analysis. Genome Biology. 2020;21:22. doi:10.1186/s13059-020-1929-3.

215. Smith JP, Dutta AB, Sathyan KM, Guertin MJ, Sheffield NC. PEPPRO: Quality control and processing of nascent RNA profiling data. Genome Biology. 2021;22:155. doi:10.1186/s13059-021-02349-4.

216. Sinha S, Satpathy AT, Zhou W, Ji H, Stratton JA, Jaffer A, et al. Profiling chromatin accessibility at single-cell resolution. Genomics, proteomics & bioinformatics. 2021;19:172–90.

217. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. Genome biology. 2016;17:63.

218. Yuan G-C, Cai L, Elowitz M, Enver T, Fan G, Guo G, et al. Challenges and emerging directions in single-cell analysis. Genome Biology. 2017;18:84. doi:10.1186/s13059-017-1218-y.

219. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016;34:525–7. doi:10.1038/nbt.3519.

220. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. Molecular cell. 2018;71:858–871.e8.

221. Baker SM, Rogerson C, Hayes A, Sharrocks AD, Rattray M. Classifying cells with scasat, a single-cell ATAC-seq analysis tool. 2019;47:e10–0. doi:10.1093/nar/gky950.

222. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biology. 2020;21:31. doi:10.1186/s13059-020-1926-6.

223. Srivastava A, Malik L, Sarkar H, Patro R. A Bayesian framework for inter-cellular information sharing improves dscRNA-seq quantification. Bioinformatics. 2020;36:i292–9. doi:10.1093/bioinformatics/btaa450.

224. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nature genetics. 2021;53:403–11. doi:10.1038/s41588-021-00790-6.

225. Denny SK, Yang D, Chuang C-H, Brady JJ, Lim JS, Grüner BM, et al. Nfib promotes metastasis through a widespread increase in chromatin accessibility. Cell. 2016;166:328–42.

226. Gomez RA, Lynch KR, Chevalier RL, Everett AD, Johns DW, Wilfong N, et al. Renin and angiotensinogen gene expression and intrarenal renin distribution during ACE inhibition. The American journal of physiology. 1988;254:F900–6.

227. Sequeira López MLS, Pentz ES, Nomasa T, Smithies O, Gomez RA. Renin cells are precursors for multiple cell types that switch to the renin phenotype when homeostasis is threatened. Developmental cell. 2004;6:719–28.

228. Martinez MF, Medrano S, Brown EA, Tufan T, Shang S, Bertoncello N, et al. Super-enhancers maintain renin-expressing cell identity and memory to preserve multi-system homeostasis. The Journal of clinical investigation. 2018;128:4787–803.

229. Gomez RA, Chevalier RL, Carey RM, Peach MJ. Molecular biology of the renal renin-angiotensin system. Kidney international Supplement. 1990;30:S18–23.

230. Gomez RA. Fate of renin cells during development and disease. Hypertension (Dallas, Tex : 1979). 2017;69:387–95.

231. Gomez RA, Sequeira-Lopez MLS. Novel functions of renin precursors in homeostasis and disease. Physiology (Bethesda, Md). 2016;31:25–33.

232. Sequeira Lopez ML, Pentz ES, Robert B, Abrahamson DR, Gomez RA. Embryonic origin and lineage of juxtaglomerular cells. American journal of physiology Renal physiology. 2001;281:F345–56.

233. Sequeira Lopez MLS, Gomez RA. Development of the renal arterioles. Journal of the American Society of Nephrology : JASN. 2011;22:2156–65.

234. Sequeira-Lopez MLS, Lin EE, Li M, Hu Y, Sigmund CD, Gomez RA. The earliest metanephric arteriolar progenitors and their role in kidney vascular development. American journal of physiology Regulatory, integrative and comparative physiology. 2015;308:R138–49.

235. Gomez RA, Sequeira-Lopez MLS. Renin cells in homeostasis, regeneration and immune defence mechanisms. Nature reviews Nephrology. 2018;14:231–45.

236. Nakamura N, Burt DW, Paul M, Dzau VJ. Negative control elements and cAMP responsive sequences in the tissue-specific expression of mouse renin genes. Proceedings of the National Academy of Sciences of the United States of America. 1989;86:56–9.

237. Horiuchi M, Nakamura N, Tang SS, Barrett G, Dzau VJ. Molecular mechanism of tissue-specific regulation of mouse renin gene expression by cAMP. Identification of an inhibitory protein that binds nuclear transcriptional factor. The Journal of biological chemistry. 1991;266:16247–54.

238. Borensztein P, Germain S, Fuchs S, Philippe J, Corvol P, Pinet F. Cis-regulatory elements and trans-acting factors directing basal and cAMP-stimulated human renin gene expression in chorionic cells. Circulation research. 1994;74:764–73.

239. Pan L, Black TA, Shi Q, Jones CA, Petrovic N, Loudon J, et al. Critical roles of a cyclic AMP responsive element and an e-box in regulation of mouse renin gene expression. The Journal of biological chemistry. 2001;276:45530–8.

240. Klar J, Sandner P, Müller MWH, Kurtz A. Cyclic AMP stimulates renin gene transcription in juxtaglomerular cells. Pflugers Archiv : European journal of physiology. 2002;444:335–44.

241. Todorov VT, Völkl S, Friedrich J, Kunz-Schughart LA, Hehlgans T, Vermeulen L, et al. Role of CREB1 and NFkappab-p65 in the down-regulation of renin gene expression by tumor necrosis factor alpha. The Journal of biological chemistry. 2005;280:24356–62.

242. Brunskill EW, Sequeira-Lopez MLS, Pentz ES, Lin E, Yu J, Aronow BJ, et al. Genes that confer the identity of the renin cell. Journal of the American Society of Nephrology : JASN. 2011;22:2213–25.

243. Castellanos Rivera RM, Monteagudo MC, Pentz ES, Glenn ST, Gross KW, Carretero O, et al. Transcriptional regulator RBP-j regulates the number and plasticity of renin cells. Physiological genomics. 2011;43:1021–8.

244. Castellanos-Rivera RM, Pentz ES, Lin E, Gross KW, Medrano S, Yu J, et al. Recombination signal binding protein for ig-kJ region regulates juxtaglomerular cell phenotype by activating the myo-endocrine program and suppressing ectopic gene expression. Journal of the American Society of Nephrology : JASN. 2015;26:67–80.

245. Lin EE, Pentz ES, Sequeira-Lopez MLS, Gomez RA. Aldo-keto reductase 1b7, a novel marker for renin cells, is regulated by cyclic AMP signaling. American journal of physiology Regulatory, integrative and comparative physiology. 2015;309:R576–84.

246. Sheffield N, Furey T. Identifying and characterizing regulatory sequences in the human genome with chromatin accessibility assays. Genes. 2012;3:651–70. doi:10.3390/genes3040651.

247. Nordström KJV, Schmidt F, Gasparoni N, Salhab A, Gasparoni G, Kattler K, et al. Unique and assay specific features of NOMe-, ATAC- and DNase i-seq data. Nucleic Acids Research. 2019;47:10580–96. doi:10.1093/nar/gkz799.

248. Chang P, Gohain M, Yen M-R, Chen P-Y. Computational methods for assessing chromatin hierarchy. 2018;16:43–53. doi:10.1016/j.csbj.2018.02.003.

249. Montefiori L, Hernandez L, Zhang Z, Gilad Y, Ober C, Crawford G, et al. Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9. 2017;7. doi:10.1038/s41598-017-02547-w.

250. Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, et al. A single-cell atlas of in vivo mammalian chromatin accessibility. Cell. 2018. doi:10.1016/j.cell.2018.06.052.

251. Rickner HD, Niu S-Y, Cheng CS. ATAC-seq assay with low mitochondrial DNA contamination from primary human CD4+ t lymphocytes. Journal of visualized experiments : JoVE. 2019.

252. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al. Practical guidelines for

the comprehensive analysis of ChIP-seq data. PLoS computational biology. 2013;9:e1003326.

253. Delisle L HF Doyle M. ATAC-seq data analysis. 2020.

254. Boyle AP, Guinney J, Crawford GE, Furey TS. F-seq: A feature density estimator for high-throughput sequence tags. Bioinformatics (Oxford, England). 2008;24:2537–8. doi:10.1093/bioinformatics/btn480.

255. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-seq (MACS). Genome biology. 2008;9:R137. doi:10.1186/gb-2008-9-9-r137.

256. McCarthy MT, O'Callaghan CA. PeaKDEck: A kernel density estimator-based peak calling program for DNaseI-seq data. Bioinformatics. 2014;30:1302–4. doi:10.1093/bioinformatics/btt774.

257. Gaspar JM. Genrich: Detecting sites of genomic enrichment. 2018.

258. Tarbell ED, Liu T. HMMRATAC: A hidden markov ModeleR for ATAC-seq. 2019;47:e91–1. doi:10.1093/nar/gkz533.

259. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25. doi:10.1186/gb-2010-11-3-r25.

260. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. Nucleic acids research. 2012;40:4288–97.

261. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15:550. doi:10.1186/s13059-014-0550-8.

262. Stark Rory BG. DiffBind: Differential binding analysis of ChIP-seq peak data. Bioconductor. 2011.

263. Tripodi IJ, Allen MA, Dowell RD. Detecting differential transcription factor activity from ATAC-seq data. Molecules (Basel, Switzerland). 2018;23.

264. Berest I, Arnold C, Reyes-Palomares A, Palla G, Rasmussen KD, Giles H, et al. Quantification of differential transcription factor activity and multiomics-based classification into activators and repressors: diffTF. Cell reports. 2019;29:3147–3159.e12.

265. Galas DJ, Schmitz A. DNAase footprinting a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 1978;5:3157.

266. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: Tools for motif discovery and searching. Nucleic acids research. 2009;37:W202–8.

267. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. 2010;38:576–89. doi:10.1016/j.molcel.2010.05.004.

268. Hashim FA, Mabrouk MS, Al-Atabany W. Review of different sequence motif finding algorithms. Avicenna journal of medical biotechnology. 2019;11:130.

269. Vierstra J, Stamatoyannopoulos JA. Genomic footprinting. Nat Methods. 2016;13:213–21. doi:10.1038/nmeth.3768.

270. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome research. 2011;21:447–55.

271. Sung M-H, Guertin MJ, Baek S, Hager GL. DNase footprint signatures are dictated by factor dynamics and DNA sequence. Molecular Cell. 2014;56:275–85. doi:10.1016/j.molcel.2014.08.016.

272. Kähärä J, Lähdesmäki H. BinDNase: A discriminatory approach for transcription factor binding prediction using DNase i hypersensitivity data. Bioinformatics (Oxford, England). 2015;31:2852–9.

273. Piper J, Elze MC, Cauchy P, Cockerill PN, Bonifer C, Ott S. Wellington: A novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic acids research. 2013;41:e201.

274. Piper J, Assi SA, Cauchy P, Ladroue C, Cockerill PN, Bonifer C, et al. Wellington-bootstrap: Differential DNase-seq footprinting identifies cell-type determining transcription factors. BMC genomics. 2015;16:1000.

275. Quach B, Furey TS. DeFCoM: Analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter. Bioinformatics (Oxford, England). 2017;33:956–63.

276. Bentsen M, Goymann P, Schultheis H, Klee K, Petrova A, Wiegandt R, et al. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nature Communications. 2020;11:4267. doi:10.1038/s41467-020-18035-1.

277. Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. Identification of transcription factor binding sites using ATAC-seq. Genome Biology. 2019;20. doi:10.1186/s13059-019-1642-2.

278. Youn A, Marquez EJ, Lawlor N, Stitzel ML, Ucar D. BiFET: Sequencing bias-free transcription factor footprint enrichment test. Nucleic acids research. 2019;47:e11.

279. Struhl K, Segal E. Determinants of nucleosome positioning. Nature Structural & Molecular Biology. 2013;20:267–73. doi:10.1038/nsmb.2506.

280. Schep AN, Buenrostro JD, Denny SK, Schwartz K, Sherlock G, Greenleaf WJ. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. Genome research. 2015;25:1757–70. doi:10.1101/gr.192294.115.

281. Vainshtein Y, Rippe K, Teif VB. NucTools: Analysis of chromatin feature occupancy profiles from high-throughput sequencing data. BMC Genomics. 2017;18:158. doi:10.1186/s12864-017-3580-2.

282. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28:495–501. doi:10.1038/nbt.1630.

283. Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics. 2015;32:289–91. doi:10.1093/bioinformatics/btv562.

284. Sheffield NC, Bock C. LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and bioconductor. Bioinformatics. 2016;32:587–9. doi:10.1093/bioinformatics/btv612.

285. Cavalcante RG, Sartor MA. Annotatr: Genomic regions in context. Bioinformatics. 2017;33:2381–3. doi:10.1093/bioinformatics/btx183.

286. Layer RM, Pedersen BS, DiSera T, Marth GT, Gertz J, Quinlan AR. GIGGLE: A search engine for large-scale integrated genome analysis. Nature Methods. 2018;15:123–6. doi:10.1038/nmeth.4556.

287. Dozmorov MG. Epigenomic annotation-based interpretation of genomic data: From enrichment analysis to machine learning. Bioinformatics. 2017;33:3323–30. doi:10.1093/bioinformatics/btx414.

288. Simovski B, Kanduri C, Gundersen S, Titov D, Domanska D, Bock C, et al. Coloc-stats: A unified web interface to perform colocalization analysis of genomic features. Nucleic Acids Research. 2018;46:W186–93. doi:10.1093/nar/gky474.

289. Boer CG de, Regev A. BROCKMAN: Deciphering variance in epigenomic regulators by k-mer factorization. BMC Bioinformatics. 2018;19. doi:10.1186/s12859-018-2255-6.

290. Fang R, Preissl S, Hou X, Lucero J, Wang X, Motamedi A, et al. Fast and accurate clustering of single cell epigenomes reveals cis-regulatory elements in rare cell types. bioRxiv. 2019;615179.

291. Yu W, Uzun Y, Zhu Q, Chen C, Tan K. scATAC-pro: A comprehensive workbench for single-cell chromatin accessibility sequencing data. Genome Biology. 2020;21:94. doi:10.1186/s13059-020-02008-0.

292. Zamanighomi M, Lin Z, Daley T, Chen X, Duren Z, Schep A, et al. Unsupervised clustering and epigenetic classification of single cells. Nature Communications. 2018;9:2410. doi:10.1038/s41467-018-04629-3.

293. Xiong L, Xu K, Tian K, Shao Y, Tang L, Gao G, et al. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. Nature Communications. 2019;10:4576. doi:10.1038/s41467-019-12630-7.

294. Ji Z, Zhou W, Ji H. Single-cell regulome data analysis by SCRAT. Bioinformatics

(Oxford, England). 2017;33:2930–2.

295. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nature methods. 2017;14:975.

296. Bravo González-Blas C, Minnoye L, Papasokrati D, Aibar S, Hulselmans G, Christiaens V, et al. cisTopic: Cis-regulatory topic modeling on single-cell ATAC-seq data. Nature Methods. 2019;16:397–400. doi:10.1038/s41592-019-0367-1.

297. Li Z, Kuppe C, Ziegler S, Cheng M, Kabgani N, Menzel S, et al. Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. Nature Communications. 2021;12:6386. doi:10.1038/s41467-021-26530-2.

298. Wang Z, Chu T, Choate LA, Danko CG. Identification of regulatory elements from nascent transcription using dREG. 2019;29:293–303. doi:10.1101/gr.238279.118.

299. Chae M, Danko CG, Kraus WL. groHMM: A computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. BMC Bioinformatics. 2015;16:222.

300. Azofeifa JG, Allen MA, Lladser ME, Dowell RD. An annotation agnostic algorithm for detecting nascent RNA transcripts in GRO-seq. IEEE/ACM transactions on computational biology and bioinformatics. 2017;14:1070–81. doi:10.1109/TCBB.2016.2520919.

301. Allison KA, Kaikkonen MU, Gaasterland T, Glass CK. Vespucci: A system for building annotated databases of nascent transcripts. Nucleic acids research. 2014;42:2433–47. doi:10.1093/nar/gkt1237.

302. Anderson WD, Duarte FM, Civelek M, Guertin MJ. Defining data-driven primary transcript annotations with primaryTranscriptAnnotation in R. Bioinformatics. 2020. doi:10.1093/bioinformatics/btaa011.

303. Sheffield NC, Stolarczyk M, Reuter VP, Rendeiro A. Linking big biomedical datasets to modular analysis with portable encapsulated projects. 2020. doi:10.1101/2020.10.08.331322.

304. Sheffield NC. Bulker: A multi-container environment manager. OSF Preprints. 2019. doi:10.31219/osf.io/natsj.

305. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature genetics. 2014;46:1311.

306. Duttke SHC, Lacadie SA, Ibrahim MM, Glass CK, Corcoran DL, Benner C, et al. Human promoters are intrinsically directional. 2015;57:674–84. doi:10.1016/j.molcel.2014.12.029.

307. Sathyan KM, McKenna BD, Anderson WD, Duarte FM, Core L, Guertin MJ. An improved auxin-inducible degron system preserves native protein levels and enables rapid and specific protein depletion. 2019;33:1441–55. doi:10.1101/gad.328237.119.

308. Andersson R, Chen Y, Core L, Lis JT, Sandelin A, Jensen TH. Human gene promoters are intrinsically bidirectional. Molecular Cell. 2015;60:346–7. doi:10.1016/j.molcel.2015.10.015.

309. Choder M, Aloni Y. RNA polymerase II allows unwinding and rewinding of the DNA and thus maintains a constant length of the transcription bubble. Journal of Biological Chemistry. 1988;263:12994–3002.

310. Shen W, Le S, Li Y, Hu F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/q file manipulation. 11:e0163962. doi:10.1371/journal.pone.0163962.

311. Martins A. fqdedup: Remove PCR duplicates from FASTQ files. 2018.

312. Daley T, Smith AD. Modeling genome coverage in single-cell sequencing. Bioinformatics. 2014;30:3159–65.

313. Rougvie AE, Lis JT. The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of d. Melanogaster is transcriptionally engaged. 1988;54:795–804. doi:10.1016/s0092-8674(88)91087-2.

314. Core LJ, Waterfall JJ, Gilchrist DA, Fargo DC, Kwak H, Adelman K, et al. Defining the status of RNA polymerase at promoters. Cell reports. 2012;2:1025–35.

315. Furey TS. ChIP–seq and beyond: New and improved methodologies to detect and characterize protein–DNA interactions. Nature Reviews Genetics. 2012;13:840–52. doi:10.1038/nrg3306.

316. Smith JP, Dutta AB, Sathyan KM, Guertin MJ, Sheffield NC. Quality control and processing of nascent RNA profiling data. Zenodo; 2021. doi:10.5281/zenodo.4542304.

317. Guertin MJ. Nascent RNA sequencing (PRO-seq) after 200nM romidepsin treatment of H9 cells. 2019.

318. Deng X, Hiatt JB, Nguyen DK, Ercan S, Sturgill D, Hillier LW, et al. Evidence for compensatory upregulation of expressed x-linked genes in mammals, caenorhabditis elegans and drosophila melanogaster. 2011;43:1179–85. doi:10.1038/ng.948.

319. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17:10–2. doi:10.14806/ej.17.1.200.

320. Oliphant TE. A guide to NumPy. Trelgol Publishing USA; 2006.

321. McKinney W, others. Data structures for statistical computing in python. In: Proceedings of the 9th python in science conference. Austin, TX; 2010. p. 51–6.

322. Stolarczyk M, Reuter VP, Magee NE, Sheffield NC. Refgenie: A reference genome resource manager. Gigascience. 2020. doi:10.1101/698704.

323. Quinlan AR. BEDTools: The swiss-army tool for genome feature analysis: BEDTools: The swiss-army tool for genome feature analysis. 2014;47:11.12.1–34. doi:10.1002/0471250953.bi1112s47.

324. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: Enabling browsing of large distributed datasets. Bioinformatics (Oxford, England). 2010;26:2204–7. doi:10.1093/bioinformatics/btq351.

325. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. 2012;9:357–9. doi:10.1038/nmeth.1923.

326. Edwards R, Edwards JA. Fastq-pair: Efficient synchronization of paired-end fastq files. BioRxiv. 2019;552885.

327. Magoc T, Salzberg SL. FLASH: Fast length adjustment of short reads to improve genome assemblies. 2011;27:2957–63. doi:10.1093/bioinformatics/btr507.

328. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. 2009;25:2078–9. doi:10.1093/bioinformatics/btp352.

329. Kurtz S, Narechania A, Stein JC, Ware D. A new method to compute k-mer frequencies and its application to annotate large repetitive plant genomes. BMC genomics. 2008;9:517.

330. Fan H, Atiya HI, Wang Y, Pisanic TR, Wang T-H, Shih I-M, et al. Epigenomic reprogramming toward mesenchymal-epithelial transition in ovarian-cancer-associated mesenchymal stem cells drives metastasis. Cell Reports. 2020;33:108473. doi:https://doi.org/10.1016/j.celrep.2020.108473.

331. Weber EW, Lynn RC, Parker KR, Anbunathan H, Lattin J, Sotillo E, et al. Transient "rest" induces functional reinvigoration and epigenetic remodeling in exhausted CAR-t cells. bioRxiv. 2020. doi:10.1101/2020.01.26.920496.

332. Zhou J, Li X, Chen J, Li T, Zhan W, Zhao J, et al. CATA: A comprehensive chromatin accessibility database for cancer. bioRxiv. 2020. doi:10.1101/2020.05.16.099325.

333. O'Connor MH, Berglind A, Kennedy MM, Keith BP, Lynch ZJ, Schaner MR, et al. BET protein inhibition regulates macrophage chromatin accessibility and microbiota-dependent colitis. bioRxiv. 2021;2021.07.15.452570. doi:10.1101/2021.07.15.452570.

334. Tovar A, Crouse WL, Smith GJ, Thomas JM, Keith BP, McFadden KM, et al. Integrative phenotypic and genomic analyses reveal strain-dependent responses to acute ozone exposure and their associations with airway macrophage transcriptional activity. bioRxiv. 2021. doi:10.1101/2021.01.29.428733.

335. Duvall E, Benitez CM, Tellez K, Enge M, Pauerstein PT, Li L, et al. Single-cell transcriptome and accessible chromatin dynamics during endocrine pancreas development. bioRxiv. 2022. doi:10.1101/2022.01.28.478217.

336. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nature. 2014;505:612–3. doi:10.1038/505612a.

337. Lauer M, Tabak L, Collins F. Opinion: The next generation researchers initiative at

NIH. Proceedings of the National Academy of Sciences of the United States of America. 2017;114:11801–3. doi:10.1073/pnas.1716941114.

338. Ram-Mohan N, Thair SA, Litzenburger UM, Cogill S, Andini N, Yang X, et al. Integrative profiling of early host chromatin accessibility responses in human neutrophils with sensitive pathogen detection. bioRxiv. 2020. doi:10.1101/2020.04.28.066829.

339. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR: An integrative and scalable software package for single-cell chromatin accessibility analysis. bioRxiv. 2020. doi:10.1101/2020.04.28.066498.

340. Fan H, Atiya H, Wang Y, Pisanic TR, Wang T-H, Shih I-M, et al. Epigenetic reprogramming towards mesenchymal-epithelial transition in ovarian cancer-associated mesenchymal stem cells drives metastasis. bioRxiv. 2020. doi:10.1101/2020.02.25.964197.

341. Anaconda software distribution. Anaconda Documentation. 2020.

342. Liu S, Li D, Lyu C, Gontarz P, Miao B, Madden P, et al. Improving ATAC-seq data analysis with AIAP, a quality control and integrative analysis package. BioRxiv. 2019;686808.

343. Pranzatelli TJ, Michael DG, Chiorini JA. ATAC2GRN: Optimized ATAC-seq and DNase1-seq pipelines for rapid and accurate genome regulatory network inference. BMC genomics. 2018;19:563.

344. Zuo Z, Jin Y, Zhang W, Lu Y, Li B, Qu K. ATAC-pipe: General analysis of genome-wide chromatin accessibility. 2019;20:1934–43. doi:10.1093/bib/bby056.

345. Sourya Bhattacharyya PV Ferhat Ay. ATACProc - a pipeline for processing ATAC-seq data. 2019.

346. Guzman C, D'Orso I. CIPHER: A flexible and extensive workflow platform for integrative next-generation sequencing data analysis and genomic regulatory element prediction. BMC bioinformatics. 2017;18:363.

347. Lee J. ENCODE ATAC-seq pipeline. 2020.

348. Wei Z, Zhang W, Fang H, Li Y, Wang X. esATAC: An easy-to-use systematic pipeline for ATAC-seq data analysis. Bioinformatics (Oxford, England). 2018. doi:10.1093/bioinformatics/bty141.

349. Divate M, Cheung E. GUAVA: A graphical user interface for the analysis and visualization of ATAC-seq data. Frontiers in genetics. 2018;9:250.

350. Ahmed Z, Ucar D. I-ATAC: Interactive pipeline for the management and pre-processing of ATAC-seq samples. PeerJ. 2017;5:e4040.

351. Ewels PA, Peltzer A, Fillinger S, Alneberg J, Patel H, Wilm A, et al. Nf-core: Community curated bioinformatics pipelines. bioRxiv. 2019;610741.

352. Tang M. pyflow-ATACseq: a snakemake based ATAC-seq pipeline. Zenodo; 2017.

doi:10.5281/zenodo.1043588.

353. Maarten van der Sande JS Siebren Frölich. seq2science. Zenodo; 2021. doi:10.5281/zenodo.4469402.

354. Bhardwaj V, Heyne S, Sikora K, Rabbani L, Rauer M, Kilpert F, et al. snakePipes: Facilitating flexible, scalable and integrative epigenomic analysis. Bioinformatics. 2019;35:4757–9.

355. Rausch T, Hsi-Yang Fritz M, Korbel JO, Benes V. Alfred: Interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long-and short-read sequencing. Bioinformatics. 2019;35:2489–91.

356. Rendeiro AF, Stolarczyk M, Reuter VP, Smith JP, Klughammer J, Schoenegger A, et al. Pypiper: A python toolkit for building restartable pipelines. 2020.

357. Stolarczyk M, Xue B, Sheffield NC. Identity and compatibility of reference genome resources. NAR Genomics and Bioinformatics. 2021;3. doi:10.1093/nargab/lqab036.

358. Amemiya HM, Kundaje A, Boyle AP. The ENCODE blacklist: Identification of problematic regions of the genome. Scientific reports. 2019;9:9354. doi:10.1038/s41598-019-45839-z.

359. Jiang H, Lei R, Ding S-W, Zhu S. Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC bioinformatics. 2014;15:182. doi:10.1186/1471-2105-15-182.

360. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for illumina sequence data. Bioinformatics (Oxford, England). 2014;30:2114–20. doi:10.1093/bioinformatics/btu170.

361. Andrews S. FastQC: A quality control tool for high throughput sequence data. 2010.

362. Wu J, Huang B, Chen H, Yin Q, Liu Y, Xiang Y, et al. The landscape of accessible chromatin in mammalian preimplantation embryos. Nature. 2016;534:652–7. doi:10.1038/nature18606.

363. Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. Journal of molecular evolution. 1994;39:174–90.

364. Richly E, Leister D. NUMTs in sequenced eukaryotic genomes. Molecular biology and evolution. 2004;21:1081–4.

365. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics (Oxford, England). 2009;25:1754–60. doi:10.1093/bioinformatics/btp324.

366. Faust GG, Hall IM. SAMBLASTER: Fast duplicate marking and structural variant read extraction. Bioinformatics (Oxford, England). 2014;30:2503–5. doi:10.1093/bioinformatics/btu314.

367. Institute B. Picard toolkit. Broad Institute, GitHub repository. 2019.

368. Koohy H, Down TA, Spivakov M, Hubbard T. A comparison of peak callers used for

DNase-seq data. PloS one. 2014;9:e96303. doi:10.1371/journal.pone.0096303.

369. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. Molecular cell. 2010;38:576–89.

370. Stolarczyk M, Reuter VP, Rendeiro AF, Smith JP, Gu A, Sheffield NC. Looper: A python-based pipeline submission engine and project manager. 2020.

371. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome research. 2012;22:1813–31. doi:10.1101/gr.136184.111.

372. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. 1990;215:403–10. doi:10.1016/s0022-2836(05)80360-2.

373. Favorov A, Mularoni L, Cope LM, Medvedeva Y, Mironov AA, Makeev VJ, et al. Exploring massive, genome scale datasets with the GenometriCorr package. PLoS Computational Biology. 2012;8:e1002529. doi:10.1371/journal.pcbi.1002529.

374. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: Archive for functional genomics data sets–update. Nucleic acids research. 2013;41:D991–5. doi:10.1093/nar/gks1193.

375. Dzau V. The cardiovascular continuum and renin-angiotensin-aldosterone system blockade. Journal of hypertension Supplement : official journal of the International Society of Hypertension. 2005;23:S9–17.

376. Rüster C, Wolf G. Angiotensin II as a morphogenic cytokine stimulating renal fibrogenesis. Journal of the American Society of Nephrology : JASN. 2011;22:1189–99.

377. Ibrahim HN, Jackson S, Connaire J, Matas A, Ney A, Najafian B, et al. Angiotensin II blockade in kidney transplant recipients. Journal of the American Society of Nephrology : JASN. 2013;24:320–7.

378. Cao W, Li A, Wang L, Zhou Z, Su Z, Bin W, et al. A salt-induced reno-cerebral reflex activates renin-angiotensin systems and promotes CKD progression. Journal of the American Society of Nephrology : JASN. 2015;26:1619–33.

379. Zhou L, Li Y, Hao S, Zhou D, Tan RJ, Nie J, et al. Multiple genes of the renin-angiotensin system are novel targets of wnt/b-catenin signaling. Journal of the American Society of Nephrology : JASN. 2015;26:107–20.

380. Kamo T, Akazawa H, Suzuki J-I, Komuro I. Roles of renin-angiotensin system and wnt pathway in aging-related phenotypes. Inflammation and regeneration. 2016;36:12.

381. Folkow B. Physiological aspects of primary hypertension. Physiological reviews. 1982;62:347–504.

382. Gomez HWAAGMAEABAXLASMASGAINALJAAMLSS-LARA. Inhibition of the renin-angiotensin system causes concentric hypertrophy of renal arterioles in mice and humans. JCI Insight. 2021;6. doi:10.1172/jci.insight.154337.

383. Remuzzi G, Bertani T. Pathophysiology of progressive nephropathies. The New England journal of medicine. 1998;339:1448–56.

384. Prasad A, Quyyumi AA. Renin-angiotensin system and angiotensin receptor blockers in the metabolic syndrome. Circulation. 2004;110:1507–12.

385. Deshayes F, Nahmias C. Angiotensin receptors: A new role in cancer? Trends in endocrinology and metabolism: TEM. 2005;16:293–9.

386. Ager EI, Neo J, Christophi C. The renin–angiotensin system and malignancy. Carcinogenesis. 2008;29:1675–84. doi:10.1093/carcin/bgn171.

387. Kamo T, Akazawa H, Komuro I. Pleiotropic effects of angiotensin II receptor signaling in cardiovascular homeostasis and aging. International heart journal. 2015;56:249–54.

388. Sobczuk P, Szczylik C, Porta C, Czarnecka AM. Renin angiotensin system deregulation as renal cancer risk factor. Oncology letters. 2017;14:5059–68.

389. Nguyen G, Delarue F, Burcklé C, Bouzhir L, Giller T, Sraer J-D. Pivotal role of the renin/prorenin receptor in angiotensin II production and cellular responses to renin. The Journal of clinical investigation. 2002;109:1417–27.

390. Santos PCJL, Krieger JE, Pereira AC. Renin-angiotensin system, hypertension, and chronic kidney disease: Pharmacogenetic implications. Journal of pharmacological sciences. 2012;120:77–88.

391. Sequeira-Lopez MLS, Gomez RA. Renin cells, the kidney, and hypertension. Circulation research. 2021;128:887–907.

392. Tan X, He W, Liu Y. Combination therapy with paricalcitol and trandolapril reduces renal fibrosis in obstructive nephropathy. Kidney international. 2009;76:1248–57.

393. Chen L, Kim SM, Eisner C, Oppermann M, Huang Y, Mizel D, et al. Stimulation of renin secretion by angiotensin II blockade is gsalpha-dependent. Journal of the American Society of Nephrology : JASN. 2010;21:986–92.

394. Guessoum O, Goes Martini A de, Sequeira-Lopez MLS, Gomez RA. Deciphering the identity of renin cells in health and disease. Trends in molecular medicine. 2021;27:280–92.

395. Petrovic N, Black TA, Fabian JR, Kane C, Jones CA, Loudon JA, et al. Role of proximal promoter elements in regulation of renin gene transcription. The Journal of biological chemistry. 1996;271:22499–505.

396. Pan L, Gross KW. Transcriptional regulation of renin: An update. Hypertension (Dallas, Tex : 1979). 2005;45:3–8.

397. Glenn ST, Jones CA, Gross KW, Pan L. Control of renin [corrected] gene expression. Pflugers Archiv : European journal of physiology. 2013;465:13–21.

398. Sequeira López MLS, Pentz ES, Nomasa T, Smithies O, Gomez RA. Renin cells are precursors for multiple cell types that switch to the renin phenotype when homeostasis is threatened. Developmental Cell. 2004;6:719–28. doi:https://doi.org/10.1016/S1534-5807(04)00134-0.

399. Surendran P, Feofanova EV, Lahrouchi N, Ntalla I, Karthikeyan S, Cook J, et al. Discovery of rare variants associated with blood pressure regulation through meta-analysis of 1.3 million individuals. Nature genetics. 2020;52:1314–32.

400. Logan CY, Nusse R. The wnt signaling pathway in development and disease. Annual review of cell and developmental biology. 2004;20:781–810.

401. Nusse R. Wnt signaling and stem cell control. Cell research. 2008;18:523–7.

402. Zhou L, Liu Y. Wnt/b-catenin signaling and renin-angiotensin system in chronic kidney disease. Current opinion in nephrology and hypertension. 2016;25:100–6.

403. Koch S. Regulation of wnt signaling by FOX transcription factors in cancer. Cancers. 2021;13. doi:10.3390/cancers13143446.

404. Wang J, Sun D, Wang Y, Ren F, Pang S, Wang D, et al. FOSL2 positively regulates TGF-β1 signalling in non-small cell lung cancer. PloS one. 2014;9:e112150.

405. Huang Y, Wongamorntham S, Kasting J, McQuillan D, Owens RT, Yu L, et al. Renin increases mesangial cell transforming growth factor-beta1 and matrix proteins through receptor-mediated, angiotensin II-independent mechanisms. Kidney international. 2006;69:105–13.

406. Dorst DCH van, Wagenaar NP de, Pluijm I van der, Roos-Hesselink JW, Essers J, Danser AHJ. Transforming growth factor-β and the renin-angiotensin system in syndromic thoracic aortic aneurysms: Implications for treatment. Cardiovascular drugs and therapy. 2021;35:1233–52.

407. Durocher M, Ander BP, Jickling G, Hamade F, Hull H, Knepp B, et al. Inflammatory, regulatory, and autophagy co-expression modules and hub genes underlie the peripheral immune response to human intracerebral hemorrhage. Journal of neuroinflammation. 2019;16:56.

408. Wang G, Lv Q, Ma C, Zhang Y, Li H, Ding Q. SMARCC1 expression is positively correlated with pathological grade and good prognosis in renal cell carcinoma. Translational andrology and urology. 2021;10:236–42.

409. Black BL, Olson EN. Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins. Annual review of cell and developmental biology. 1998;14:167–96.

410. Naya FJ, Olson E. MEF2: A transcriptional target for signaling pathways controlling skeletal muscle growth and differentiation. Current opinion in cell biology. 1999;11:683–8.

411. Perry R. L. S, Yang C, Soora N, Salma J, Marback M, Naghibi L, et al. Direct interaction between myocyte enhancer factor 2 (MEF2) and protein phosphatase 1α represses MEF2-dependent gene expression. Molecular and Cellular Biology. 2009;29:3355–66. doi:10.1128/MCB.00227-08.

412. Estrella NL, Desjardins CA, Nocco SE, Clark AL, Maksimenko Y, Naya FJ. MEF2 transcription factors regulate distinct gene programs in mammalian skeletal muscle differentiation*. Journal of Biological Chemistry. 2015;290:1256–68. doi:https://doi.org/10.1074/jbc.M114.589838.

413. Pon JR, Marra MA. MEF2 transcription factors: Developmental regulators and emerging cancer genes. Oncotarget. 2016;7:2297–312.

414. Kato Y, Kravchenko VV, Tapping RI, Han J, Ulevitch RJ, Lee JD. BMK1/ERK5 regulates serum-induced early gene expression through transcription factor MEF2C. The EMBO journal. 1997;16:7054–66.

415. Sacilotto N, Chouliaras KM, Nikitenko LL, Lu YW, Fritzsche M, Wallace MD, et al. MEF2 transcription factors are key regulators of sprouting angiogenesis. Genes & development. 2016;30:2297–309.

416. Tamahara T, Ochiai K, Muto A, Kato Y, Sax N, Matsumoto M, et al. The mTOR-Bach2 cascade controls cell cycle and class switch recombination during b cell differentiation. Molecular and cellular biology. 2017;37.

417. Miura Y, Morooka M, Sax N, Roychoudhuri R, Itoh-Nakadai A, Brydun A, et al. Bach2 promotes b cell receptor–induced proliferation of b lymphocytes and represses cyclin-dependent kinase inhibitors. The Journal of Immunology. 2018;200:2882–93. doi:10.4049/jimmunol.1601863.

418. Igarashi K, Kurosaki T, Roychoudhuri R. BACH transcription factors in innate and adaptive immunity. Nature reviews Immunology. 2017;17:437–50.

419. Jang E, Kim UK, Jang K, Song YS, Cha J-Y, Yi H, et al. Bach2 deficiency leads autoreactive b cells to produce IgG autoantibodies and induce lupus through a t cell-dependent extrafollicular pathway. Experimental & Molecular Medicine. 2019;51:1–3. doi:10.1038/s12276-019-0352-x.

420. Kurokawa H, Motohashi H, Sueno S, Kimura M, Takagawa H, Kanno Y, et al. Structural basis of alternative DNA recognition by maf transcription factors. Molecular and cellular biology. 2009;29:6232–44.

421. Germain S, Konoshita T, Philippe J, Corvol P, Pinet F. Transcriptional induction of the human renin gene by cyclic AMP requires cyclic AMP response element-binding protein (CREB) and a factor binding a pituitary-specific trans-acting factor (Pit-1) motif. Biochemical

Journal. 1996;316:107–13. doi:10.1042/bj3160107.

422. Yao C, Lou G, Sun H-W, Zhu Z, Sun Y, Chen Z, et al. BACH2 enforces the transcriptional and epigenetic programs of stem-like CD8(+) t cells. Nature immunology. 2021;22:370–80.

423. Belyea BC, Xu F, Pentz ES, Medrano S, Li M, Hu Y, et al. Identification of renin progenitors in the mouse bone marrow that give rise to b-cell leukaemia. Nature communications. 2014;5:3273.

424. Lin EE, Sequeira-Lopez MLS, Gomez RA. RBP-j in FOXD1+ renal stromal progenitors is crucial for the proper development and assembly of the kidney vasculature and glomerular mesangial cells. American journal of physiology Renal physiology. 2014;306:F249–58.

425. Magaletta ME, Lobo M, Kernfeld EM, Aliee H, Huey JD, Parsons TJ, et al. Integration of single-cell transcriptomes and chromatin landscapes reveals regulatory programs driving pharyngeal organ development. Nature Communications. 2022;13:457. doi:10.1038/s41467-022-28067-4.

426. Marand AP, Chen Z, Gallavotti A, Schmitz RJ. A cis-regulatory atlas in maize at single-cell resolution. Cell. 2021;184:3041–3055.e21.

427. Ma Z, Lytle NK, Ramos C, Naeem RF, Wahl GM. Single-cell transcriptomic and epigenetic analyses of mouse mammary development starting with the embryo. In: Vivanco MdM, editor. Mammary stem cells: Methods and protocols. New York, NY: Springer US; 2022. p. 49–82. doi:10.1007/978-1-0716-2193-6_3.

428. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral t cell exhaustion. Nature biotechnology. 2019;37:925–36.

429. Granja JM, Klemm S, McGinnis LM, Kathiria AS, Mezger A, Corces MR, et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nature biotechnology. 2019;37:1458–65.

430. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with harmony. Nature Methods. 2019;16:1289–96. doi:10.1038/s41592-019-0619-0.

431. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England). 2013;29:15–21.

432. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM3rd, et al. Comprehensive integration of single-cell data. Cell. 2019;177:1888–1902.e21.

433. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in r. Bioinformatics (Oxford, England). 2017;33:1179–86.

434. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative analysis of single-cell RNA sequencing methods. Molecular cell. 2017;65:631–643.e4.

435. Osorio D, Cai JJ. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. Bioinformatics (Oxford, England). 2021;37:963–7.

436. Mercer TR, Neph S, Dinger ME, Crawford J, Smith MA, Shearwood A-MJ, et al. The human mitochondrial transcriptome. Cell. 2011;146:645–58.

437. AlJanahi AA, Danielsen M, Dunbar CE. An introduction to the analysis of single-cell RNA-sequencing data. Molecular therapy Methods & clinical development. 2018;10:189–96.

438. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome biology. 2019;20:296.

439. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. 2015;161:1202–14. doi:10.1016/j.cell.2015.05.002.

440. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. 2018;361:1380–5. doi:10.1126/science.aau0730.

441. Park J, Shrestha R, Qiu C, Kondo A, Huang S, Werth M, et al. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. Science (New York, NY). 2018;360:758–63.

442. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014;158:1431–43.

443. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nature methods. 2017;14:975–8.

444. Grant CE, Bailey TL, Noble WS. FIMO: Scanning for occurrences of a given motif. Bioinformatics (Oxford, England). 2011;27:1017–8.

445. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Research. 2021;50:D165–73. doi:10.1093/nar/gkab1113.

446. Lawrence WAP Michael AND Huber. Software for computing and annotating genomic ranges. PLOS Computational Biology. 2013;9:1–0. doi:10.1371/journal.pcbi.1003118.

447. Sheffield NC, Stolarczyk M, Reuter VP, Rendeiro AF. Linking big biomedical datasets to modular analysis with Portable Encapsulated Projects. GigaScience. 2021;10.

doi:10.1093/gigascience/giab077.

448. Stolarczyk M, Reuter VP, Smith JP, Magee NE, Sheffield NC. Refgenie: a reference genome resource manager. GigaScience. 2020;9. doi:10.1093/gigascience/giz149.

449. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921. doi:10.1038/35057062.

450. Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420:520–62. doi:10.1038/nature01262.

451. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic acids research. 2016;44:D733–45.

452. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic Acids Research. 2020;49:D884–91. doi:10.1093/nar/gkaa942.

453. Lee BT, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, et al. The UCSC genome browser database: 2022 update. Nucleic acids research. 2022;50:D1115–22.

454. Gharavi E, Gu A, Zheng G, Smith JP, Zhang A, Brown DE, et al. Embeddings of genomic region sets capture rich biological associations in low dimensions. Bioinformatics. 2021. doi:10.1093/bioinformatics/btab439.

455. Chen H, Lareau C, Andreani T, Vinyard ME, Garcia SP, Clement K, et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. 2019;20. doi:10.1186/s13059-019-1854-5.

456. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing [review article]. IEEE Computational Intelligence Magazine. 2018;13:55–75. doi:10.1109/MCI.2018.2840738.

457. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in neural information processing systems. Curran Associates, Inc.; 2013.

458. Kupkova K SJ Mosquera JV. GenomicDistributions: Fast analysis of genomic intervals with bioconductor. BMC Genomics. 2022. doi:10.1186/s12864-022-08467-y.

459. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis. 2010;31:27–36.

460. Baylin SB, Jones PA. Epigenetic determinants of cancer. Cold Spring Harbor perspectives in biology. 2016;8.

461. Zoghbi HY, Beaudet AL. Epigenetics and human disease. Cold Spring Harbor perspectives in biology. 2016;8:a019497.

462. Tzika E, Dreker T, Imhof A. Epigenetics and metabolism in health and disease. Frontiers in Genetics. 2018;9. doi:10.3389/fgene.2018.00361.

463. Mazzone R, Zwergel C, Artico M, Taurone S, Ralli M, Greco A, et al. The emerging role of epigenetics in human autoimmune disorders. Clinical Epigenetics. 2019;11:34. doi:10.1186/s13148-019-0632-2.

464. github. GitHub. 2020.

465. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, et al. HiChIP: Efficient and sensitive analysis of protein-directed genome architecture. Nature Methods. 2016;13:919–22. doi:10.1038/nmeth.3999.

466. Ford ES. Trends in mortality from all causes and cardiovascular disease among hypertensive and nonhypertensive adults in the united states. Circulation. 2011;123:1737–44.

467. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: A report of the american college of cardiology/american heart association task force on clinical practice guidelines. Journal of the American College of Cardiology. 2018;71:e127–248. doi:https://doi.org/10.1016/j.jacc.2017.11.006.

468. Baxter SM, Day SW, Fetrow JS, Reisinger SJ. Scientific software development is not an oxymoron. PLoS computational biology. 2006;2:e87.

469. Hannay JE, MacLeod C, Singer J, Langtangen HP, Pfahl D, Wilson G. How do scientists develop and use scientific software? In: 2009 ICSE workshop on software engineering for computational science and engineering. 2009. p. 1–8. doi:10.1109/SECSE.2009.5069155.

470. Merali Z. Computational science: ...error. Nature. 2010;467:775–7.

471. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, et al. Best practices for scientific computing. PLoS biology. 2014;12:e1001745.

472. Lawlor B, Walsh P. Engineering bioinformatics: Building reliability, performance and productivity into bioinformatics software. Bioengineered. 2015;6:193–203.

473. List M, Ebert P, Albrecht F. Ten simple rules for developing usable software in computational biology. In: PLoS computational biology. 2017. p. e1005265.

474. Zappia L, Phipson B, Oshlack A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS computational biology. 2018;14:e1006245.