nature research

https://doi.org/10.1038/s43586-020-00008-9

Supplementary information

Chromatin accessibility profiling methods

In the format provided by the authors and unedited

Supplementary Table 1. Commonly used bioinformatics tools for data processing and analysis of bulk and single-cell chromatin accessibility data. For each tool, the processing step during which it is used, specific features and programming language are given. All mentioned tools are open source and discussed in the text. Unless specified in the 'Specific features and utilities' column, the tools are not exclusive to a specific type of chromatin accessibility data.

Processing step	Tool	Specific features and utilities	Language	Reference
Sequencing data quality control	FastQC	Checks sequence quality, GC content, sequence duplication levels, overrepresented sequences etc.	Java	[1]
Trimming of sequencing adapters	cutadapt	Finds and removes unwanted adapter sequences, primers, poly-A tails from sequencing reads.	Python	[2]
	trimmomatic	Performs read trimming and filtering, including the identification of adapter sequences and quality filtering.	Java	[3]
	fastq-mcf	Detects and removes sequencing adapters. Includes other filtering steps, like filtering of too short reads after trimming.	C++	[4]
Mapping	Bowtie2	Read aligner, used in the ENCODE ATAC-seq pipeline.	C++	[5]
	bwa-mem	Read aligner, used in the Cell Ranger scATAC-seq pipeline.	C	[6]
Peak calling	MACS2	Peak caller originally designed for ChIP-seq data, but also the most commonly used peak caller for chromatin accessibility data.	Python	[7]
	ZINBA	Peak caller which identifies enriched sites across a variety of omics datasets.	C, R	[8]
	HMMRATAC	Peak caller specific for ATAC-seq data, based on a semi-supervised machine learning approach.	Java	[9]
	Genrich	Peak caller with specific analysis modes for ChIP-seq and ATAC-seq data, also includes optional PCR duplicate removal.	C	[10]
	F-seq	Peak caller which generates a continuous tag sequence density estimation that allows the identification of peaks in DNase-seq and ATAC-seq data.	Java	[11]
	Hotspot	Peak caller for DNase-seq and ATAC-seq data based on the binomial distribution using a local background model.	C++	[12]
	DANPOS	Performs dynamic nucleosome analysis at single-nucleotide resolution in MNase-seq data.	Python, R	[13]

Nucleosome positioning tools	NucleoATAC	Calls nucleosome positions and occupancy using ATAC-seq data.	Python	[14]
Consensus peak calling	consensusSeekeR	Compares genomic positions and ranges from multiple experiments to extract common regions.	R	[15]
	DiffBind	Outputs consensus peaks by merging peaks from all input peak sets.	R	[16]
Differential accessibility	HOMER	Uses findPeaks or getDifferentialPeaks to identify peaks that are significantly enriched in an experiment sample versus a control/background sample.	Perl and C++	[17]
	DESeq2	Originally designed for differential gene expression, but can perform a differential accessibility analysis when providing a non-normalized count matrix of peaks versus samples.	R	[18]
	MACS2	Calls differential peaks when providing an experiment sample (as -t) and control sample (as -c), also replicates can be included in the comparison.	Python	[19]
	DiffBind	Computes differentially bound sites from multiple experiments using quantitative data.	R	[20]
	DBChIP	Detects differentially bound sharp binding sites in ChIP-seq data mostly.	R	[21]
Clustering	deepTools	Suite containing several tools, including tools to perform clustering and generate heatmaps of count matrices, allowing the identification of differential region clusters.	Python	[22]
Region set enrichment analysis	GREAT	Assigns biological enrichments to a set of regions by analyzing the annotations of the nearby genes.	C, available as web tool	[23]
	ChIPseeker	Retrieves nearest genes around peaks, annotates regions, and calculates overlap with ChIP-seq datasets.	R	[24]
	ChIPpeakAnno	Includes functions to obtain Gene Ontology terms and to find nearest genes around peaks.	R	[25]
	Enrichr	Converts peaks to genes and performs gene list enrichment analysis.	Python, available as web tool and in R	[26]
	cisTarget	Identifies significant enrichment of public DHS, FAIRE and ChIP-seq data within sets of regions.	Web tool, also available in R, Python	[27]
	GIGGLE	Generic search engine for statistically measuring the overlap of a set of regions with any genomic annotation.	C, Python	[28]

	LOLA	Enrichment analysis of region sets against public databases and custom collections of annotated genomic region sets.	R	[29]
Chromatin segmentation	ChromHMM	Software for learning and characterizing chromatin states based on a multivariate Hidden Markov Model.	Java	[30]
	EpicSeg	Segmentation and characterization of epigenomic landscapes by using a probabilistic model.	R	[31]
	Segway	Performs semi-automated genome annotation, based on a dynamic Bayesian network model.	Python	[32]
Motif analysis	HOMER	Performs motif discovery via findMotifs and findMotifsGenome to (de novo) discover enriched motifs in a set of regions compared to a background. A user-provided control set can also be used as background.	Perl and C++	[17]
	MEME	Contains several motif-based sequence analysis tools, performs motif scanning or de novo motif discovery.	Python, available as web tool	[33]
	cisTarget	Motif enrichment analysis in a set of genomic regions using a ranking-and-recovery procedure.	Web tool, also available in R, Python	[27]
	RSAT	Motif analysis and de novo motif discovery within a set of regulatory regions, including spaced pairs of k-mers.	Web tool, also available for command line	[34]
	Weeder	Discovers TF binding sites.	C	[35]
Deep learning	DeepLIFT	Deep learning model that allows to unravel the output prediction of a neural network by backpropagation, revealing informative nucleotides within sequences.	Python	[36]
	Basset	Deep learning model that models peak-based chromatin profiles, can be applied to neural networks to learn the importance of nucleotides within DNA sequences	Python, also provided as Jupyter notebooks	[37]
	DeapSEA	Deep learning model that can be trained on chromatin accessibility data and that predicts the importance of nucleotide alterations.	Python, available as web tool and Jupyter notebooks	[38]
TF footprinting	Wellington algorithm	Algorithm that identifies protein-DNA binding sites from DNase-seq data using a binomial test. Implemented in the pyDNase package.	Python	[39]
	HINT(-ATAC)	Hidden Markov model-based method that identifies footprints in ATAC-seq or DNase-seq data.	Python	[40, 41]
	TOBIAS	TF footprinting framework for ATAC-seq data. Includes a Tn5 bias correction module that utilizes a	Python	[42]

		dinucleotide weight matrix to estimate the background bias of the Tn5 transposase.		
	DBFP	Dynamic Bayesian network-based approach to identify protein-binding footprints.	Python	[43]
	DNase2TF	Identifies footprint candidates in DNase-seq data by assessing the significance of cut depletion compared to a local background using a z-score.	R, Perl	[44]
	CENTIPEDE	Hierarchical Bayesian mixture model, infers regions in the genome bound by TFs in DNase-seq data.	R	[45]
	FLR	Builds bias corrected and TF-specific footprint models and uses a log-likelihood ratio to define footprint scores from DNase-seq data.	R	[46]
	DeFCoM	Support vector machine-based supervised learning footprinter, used to perform footprint classification. Footprinter for DNase-seq and ATAC-seq data.	Python	[47]
Genome browsers	IGV	Allows interactive visual exploration of genomic data, can be downloaded as a desktop application.	Java, web tool	[48]
	UCSC Genome Browser	Allows interactive visual exploration of genomic data, allows easy loading of public data such as conservation tracks, SNPs etc.	Web tool, C, can be mirrored locally	[49]
	Ensembl	Provides a genome browser that acts as a single point of access to annotated genomes for mainly vertebrate species.	Web tool	[50]
	JBrowse	Allows interactive visual exploration of genomic data, can be easily embedded into websites.	JavaScript	[51, 52]
scATAC analysis tools	chromVar	scATAC analysis tool that includes and is based on the identification of motifs with variability in chromatin accessibility between cells or samples.	R	[53]
	cisTopic	scATAC analysis tool that is based on an unsupervised Bayesian framework using latent Dirichlet allocation, allowing simultaneous clustering of single cells and regions.	R	[54]
	scABC	scATAC analysis tool based on a weighted K-medoids clustering for unsupervised clustering and analysis of scATAC-seq data.	R	[55]
	Scasat	Pipeline that combines both the initial processing of scATAC-seq data and further downstream analyses.	Python, R, provided as Jupyter notebooks	[56]
	SnapATAC	scATAC analysis tool based on count features within fixed-size genomic bins. Dissects cellular heterogeneity in an unbiased manner and maps the trajectories of cellular states. Allows analysis of large datasets.	Python, R	[57]

	ArchR	Extensive suite for fast scATAC-seq data processing and analysis.	R	[58]
	BROCKMAN	scATAC analysis tool that uses k-mer factorization to infer variation in TF activity across samples.	Bash and R	[59]
	SCRAT	scATAC analysis tool based on summarized features across a feature database. Includes a graphical interface.	Web tool, R	[60]
	Cicero	scATAC analysis tool. In addition, allows for the prediction of cis-regulatory interactions by examining co-accessibility, and for trajectory inference (based on Monocle).	R	[61]
	Signac	Extension of the scRNA-seq analysis toolkit Seurat for the analysis, interpretation and exploration of single-cell chromatin accessibility data.	R	[62]
	MAESTRO	Comprehensive scRNA-seq and scATAC-seq analysis suit, integrates dozens of other tools and packages in a single pipeline.	C	[63]
	EpiScanpy	Extension of the scRNA-seq analysis toolkit Scanpy for the analysis of scATAC-seq and single-cell DNA methylation data	Python, also provided as Jupyter notebooks	[64]
Batch effect removal in single-cell data	BBKNN	Performs batch effect removal through a k nearest-neighbor approach.	Python	[65]
	Scanorama	Enables batch correction based on mutual nearest-neighbors matching.	Python, R	[66]
	scVI	Package for single-cell omics data, which includes a generative model that allows for batch correction.	Python	[67]
Single-cell trajectory inference	Cicero	scATAC-seq analysis tool which includes trajectory reconstruction via implementing a modification of the scRNA-seq trajectory inference tool Monocle	R	[61]
	STREAM	Performs trajectory inference for scRNA-seq and scATAC-seq data based on an Elastic Principle Graph Implementation (ElPiGraph)	Python, web tool	[68]

Supplementary References

1. Babraham Bioinformatics. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

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

3. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* **30**, 2114–2120 (2014).

4. Aronesty, E. *ea-utils : 'Command-line tools for processing biological sequencing data'*. (2011).

5. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

6. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).

7. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

8. Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W. & Lieb, J. D. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol.* **12**, R67 (2011).

9. Tarbell, E. D. & Liu, T. HMMRATAC: a Hidden Markov ModeleR for ATAC-seq. *Nucleic Acids Res.* **47**, e91–e91 (2019).

10. Gaspar, J. M. *Genrich: detecting sites of genomic enrichment*. Github, https://github.com/jsh58/Genrich (2018).

11. Boyle, A. P., Guinney, J., Crawford, G. E. & Furey, T. S. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinforma. Oxf. Engl.* **24**, 2537–2538 (2008).

12. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).

13. Chen, K. *et al.* DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res.* **23**, 341–351 (2013).

14. Schep, A. N. *et al.* Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.* **25**, 1757–1770 (2015).

15. Samb, R. *et al.* Using informative Multinomial-Dirichlet prior in a t-mixture with reversible jump estimation of nucleosome positions for genome-wide profiling. *Stat. Appl. Genet. Mol. Biol.* **14**, (2015).

16. Stark, R. & Brown, G. *DiffBind: differential binding analysis of ChIP-Seq peak data.* (2011).

17. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

18. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

19. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

20. Stark, R. & Brown, G. *DiffBind: differential binding analysis of ChIP-Seq peak data.* (2011).

21. Liang, K. & Keles, S. Detecting differential binding of transcription factors with ChIP-seq. *Bioinforma. Oxf. Engl.* **28**, 121–122 (2012).

22. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-165 (2016).

23. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).

24. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).

25. Zhu, L. J. *et al.* ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11**, 237 (2010).

26. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

27. Imrichová, H., Hulselmans, G., Kalender Atak, Z., Potier, D. & Aerts, S. i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Res.* **43**, W57–W64 (2015).

28. Layer, R. M. *et al.* GIGGLE: a search engine for large-scale integrated genome analysis. *Nat. Methods* **15**, 123–126 (2018).

29. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinforma. Oxf. Engl.* **32**, 587–589 (2016).

30. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).

31. Mammana, A. & Chung, H.-R. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biol.* **16**, 151 (2015).

32. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012).

33. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).

34. Thomas-Chollier, M. *et al.* RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res.* **40**, e31–e31 (2012).

35. Pavesi, G., Mereghetti, P., Mauri, G. & Pesole, G. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* **32**, W199–W203 (2004).

36. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. *ArXiv170402685 Cs* (2017).

37. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).

38. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–4 (2015).

39. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41**, e201 (2013).

40. Gusmao, E. G., Dieterich, C., Zenke, M. & Costa, I. G. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics* **30**, 3143–3151 (2014).

41. Li, Z. *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome Biol.* **20**, 45 (2019).

42. Bentsen, M. *et al.* ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat. Commun.* **11**, 4267 (2020).

43. Chen, X., Hoffman, M. M., Bilmes, J. A., Hesselberth, J. R. & Noble, W. S. A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics* **26**, i334–i342 (2010).

44. Sung, M.-H., Guertin, M. J., Baek, S. & Hager, G. L. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol. Cell* **56**, 275–285 (2014).

45. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21**, 447–455 (2011).

46. Yardımcı, G. G., Frank, C. L., Crawford, G. E. & Ohler, U. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res.* **42**, 11865–11878 (2014).

47. Quach, B. & Furey, T. S. DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter. *Bioinforma. Oxf. Engl.* **33**, 956–963 (2017).

48. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

49. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

50. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res.* gkz966 (2019) doi:10.1093/nar/gkz966.

51. Buels, R. *et al.* JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* **17**, 66 (2016).

52. Hofmeister, B. T. & Schmitz, R. J. Enhanced JBrowse plugins for epigenomics data visualization. *BMC Bioinformatics* **19**, 159 (2018).

53. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, (2017).

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

55. Zamanighomi, M. *et al.* Unsupervised clustering and epigenetic classification of single cells. *Nat. Commun.* **9**, (2018).

56. Baker, S. M., Rogerson, C., Hayes, A. & Sharrocks, A. D. Classifying cells with Scasat - a tool to analyse single-cell ATAC-seq. 1–16 (2017).

57. Fang, R. *et al. SnapATAC: A Comprehensive Analysis Package for Single Cell ATAC-seq*. http://biorxiv.org/lookup/doi/10.1101/615179 (2019) doi:10.1101/615179.

58. Granja, J. M. *et al. ArchR: An integrative and scalable software package for single-cell chromatin accessibility analysis*. http://biorxiv.org/lookup/doi/10.1101/2020.04.28.066498 (2020) doi:10.1101/2020.04.28.066498.

59. de Boer, C. G. & Regev, A. BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization. *BMC Bioinformatics* **19**, (2018).

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

61. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* **71**, 858-871.e8 (2018).

62. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

63. Wang, C. *et al.* Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol.* **21**, 198 (2020).

64. Danese, A., Richter, M. L., Fischer, D. S., Theis, F. J. & Colomé-Tatché, M. *EpiScanpy: integrated single-cell epigenomic analysis*. http://biorxiv.org/lookup/doi/10.1101/648097 (2019) doi:10.1101/648097.

65. Polański, K. *et al.* BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* btz625 (2019) doi:10.1093/bioinformatics/btz625.

66. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).

67. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).

68. Chen, H. *et al.* Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat. Commun.* **10**, 1903 (2019).