METHOD ARTICLE

Fast analysis of scATAC-seq data using a predefined set of genomic regions [version 1; peer review: awaiting peer review]

Valentina Giansanti [1,2], Ming Tang[3], Davide Cittaro [2]

[1]Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy
[2]Center for Omics Sciences, IRCCS San Raffaele Institute, Milan, Italy
[3]FAS informatics, Harvard University, Cambridge, MA, USA

v1 **First published:** 20 Mar 2020, **9**:199 (https://doi.org/10.12688/f1000research.22731.1)
Latest published: 20 Mar 2020, **9**:199 (https://doi.org/10.12688/f1000research.22731.1)

Abstract

Background: Analysis of scATAC-seq data has been recently scaled to thousands of cells. While processing of other types of single cell data was boosted by the implementation of alignment-free techniques, pipelines available to process scATAC-seq data still require large computational resources. We propose here an approach based on pseudoalignment, which reduces the execution times and hardware needs at little cost for precision.
Methods: Public data for 10k PBMC were downloaded from 10x Genomics web site. Reads were aligned to various references derived from DNase I Hypersensitive Sites (DHS) using *kallisto* and quantified with *bustools*. We compared our results with the ones publicly available derived by *cellranger-atac*.
Results: We found that *kallisto* does not introduce biases in quantification of known peaks and cells groups are identified in a consistent way. We also found that cell identification is robust when analysis is performed using DHS-derived reference in place of *de novo* identification of ATAC peaks. Lastly, we found that our approach is suitable for reliable quantification of gene activity based on scATAC-seq signal, thus allows for efficient labelling of cell groups based on marker genes.
Conclusions: Analysis of scATAC-seq data by means of *kallisto* produces results in line with standard pipelines while being considerably faster; using a set of known DHS sites as reference does not affect the ability to characterize the cell populations

Keywords

single cell, scATAC-seq, pseudoalignment

Open Peer Review

Reviewer Status *AWAITING PEER REVIEW*

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Davide Cittaro (cittaro.davide@hsr.it)

Author roles: Giansanti V: Data Curation, Formal Analysis, Resources, Software, Writing – Original Draft Preparation; **Tang M**: Data Curation, Formal Analysis, Resources, Software, Writing – Original Draft Preparation; **Cittaro D**: Conceptualization, Investigation, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: DC and VG are supported by the Accelerator Award: A26815 entitled: "Single-cell cancer evolution in the clinic" funded through a partnership between Cancer Research UK and Fondazione AIRC. MT is supported by NIH grants 1U19MH114830 and 1U19MH114821.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Giansanti V, Tang M and Cittaro D. **Fast analysis of scATAC-seq data using a predefined set of genomic regions [version 1; peer review: awaiting peer review]** F1000Research 2020, **9**:199 (https://doi.org/10.12688/f1000research.22731.1)

First published: 20 Mar 2020, **9**:199 (https://doi.org/10.12688/f1000research.22731.1)

Introduction

Recent technological advances in single-cell technologies resulted in a tremendous increase in the throughput in a relatively short span of time[1]. The increasing number of cells that could be analyzed prompted a better usage of computational resources; this has been especially true for the post-alignment and quantification phases. As a consequence, it is today feasible to run the analysis of single cell data on commodity hardware with limited resources[2], even when the number of observables is in the order of hundreds of thousands. Conversely, the analysis steps from raw sequences to count matrices lagged for some time; alignment to the reference genome or transcriptome is largely dependent on classic aligners, without any specific option to handle single-cell data, with the notable exception of the latest implementation of STARsolo in the STAR aligner[3].

More recently, analysis of NGS data benefit from technologies based on *k*-mer processing, allowing alignment-free sequence comparison[4]. Most of these technologies require a catalog of *k*-mers expected to be in the dataset and, hence, subject of quantification. RNA-seq analysis relies on the quantification of gene/transcript abundances and, while it is possible to perform *de novo* characterization of unknown species in every experiment, it is common practice[5,6] to rely on a well-defined gene model such as GENCODE[7] to quantify expressed species. It is then possible to efficiently perform alignment-free analysis on transcripts to quantify gene abundances and, in fact, tools implementing this approach such as *kallisto*[8] or Salmon[9] have been quickly adopted on a wide scale. Moreover, a recent implementation of *kallisto* extended its capabilities to the analysis of single cell RNA-seq data[10] by direct handling of cell barcodes and UMIs, allowing the analysis of such data in a streamlined way.

Analysis of epigenetic features by ATAC-seq requires the identification of enriched peaks along the genome sequence. This is typically achieved using peak callers such as MACS[11], opportunely tuned. Since ATAC-seq signal mirrors DNA accessibility as mapped by DNase-seq assays[12] and catalogs of DNase I Hypersensitive Sites (DHS) are available[13,14] it should be possible to perform reference-based ATAC-seq analysis in a way much similar to what is performed for RNA-seq analysis. In this paper we show it is indeed possible to perform single-cell ATAC-seq analysis using *kallisto* and *bustools*, with minor tweaks, using an indexed reference of ~1 million known DHS sites on the human genome.

Methods

Single cell ATAC-seq data

Single cell ATAC-seq data were downloaded from the 10x Genomics public datasets (https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_pbmc_10k) and include sequences for 10*k* PBMC from a healthy donor. We used the Peak by cell matrix HDF5 (filtered) object as our ground truth.

Generation of *kallisto* index

We downloaded the DNase I Hypersensitive Sites (DHS) interval list for hg19 genome from the Regulatory Elements DB[15], intervals closer than 500*bp* were clustered using bedtools[16].

We extracted DNA sequences for DHS intervals and indexed corresponding fasta files using *kallisto index* (v0.46.0) with default parameters, resulting in an index for the full DHS set (iDHSfull) and an index for the merged set (iDHS500). The same procedure was performed for the peak set identified by *cellranger-atac* and distributed along with the data (iMACS).

Peak quantification

kallisto requires the definition of the unique molecular identifiers (UMI) and cellular barcodes (CB) in a specific fastq file. For standard Chromium scRNA-seq data, these are substrings of R1 and RNA is sequenced in R2. Chromium scATAC-seq reads are not structured in the same way, paired end genomic reads are in R1 and R3, R2 includes only the 16*bp* cellular barcode. In addition, `kallisto bus` expects only a single read with genomic information. Therefore we simulated appropriate structures in three different ways:

1. by adding 12 random nucleotides and mapping the R1 file (forward read):
```
kallisto bus -x 10xV2 modified_R1.fastq.gz
pbmc_10k_R1.fastq.gz
```
2. by extracting sequences of different length *n* (5, 10, 15, 20) from the 5' of R3 (reverse read) and mapping the R1 file:
```
kallisto bus -x 1,0,16:2,0,n:0,0,0
pbmc_10k_R1.fastq.gz
pbmc_10k_R2.fastq.gz
pbmc_10k_R3.fastq.gz
```
3. by extracting sequences of different length *n* (5, 10, 15, 20) from the 5' of R1 and then mapping the R3 file:
```
kallisto bus -x 1,0,16:2,0,n:0,0,0
pbmc_10k_R3.fastq.gz
pbmc_10k_R2.fastq.gz
pbmc_10k_R1.fastq.gz
```

We will refer to the second set of simulation as *n-fwd* and to the third set as *n-rev*, where *n* is the number of nucleotides considered as UMI. We also applied two different summarization strategies for `bustools count` Step. In the first approach, pseudocounts are not summarized, the number of features matches the size of the index; in the second approach, summarized, we let `bustools map` counts on iDHSfull to the merged intervals (Figure 1A).

Analsyis of single-cell data

Counts matrices were analysed using Scanpy (v1.4.2)[2] with standard parameters. We filtered out cells that had less than 200 regions and regions that were not at least in 10 cells. The count matrices were normalized and log transformed. The highly variable regions were selected and the subsetted matrices processed to finally clusterized the data with the Leiden algorithm[17]. Adjusted mutual information (MI) was used to evaluate the concordance between the 10x and our matrices.

The matrices derived from *kallisto* and *cellranger-atac* were also imported into Seurat V3[18]. Gene activity score was calculated using the CreateGeneActivityMatrix function or directly summarized

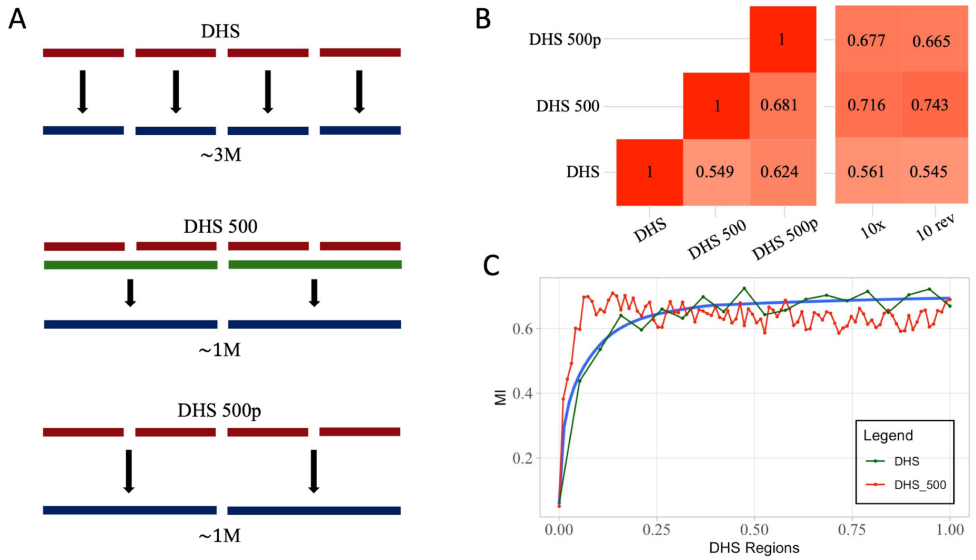

Figure 1. (**A**) Graphical depiction of processing of pseudoalignment over DHS, based on three DHS derived indices. The first (DHS) generated by *kallisto* on *~2M* DNase I sites, the second (DHS500) by merging regions closer than 500*bp* and the last (DHS500p) by projecting the result of DHS index to DHS500 using bustools capabilities. (**B**) Heatmaps representing MI scores for the DHS derived matrices. The heatmap on the left reports the pairwise MI values between DHS, DHS500 and DHS500p strategies. The heatmap on the right represents MI values comparing the DHS derived strategies to the *cellranger-atac* (10x) results or 10-*rev* strategy. DHS500 strategy achieves the highest scores. (**C**) MI values comparing DHS (green line) and DHS500 (red line) strategies to *cellranger-atac* at different thresholds on the number of regions considered in the analysis. When approximately 50, 000 regions are included, the MI stabilizes at its maximum.

by *kallisto*. The annotated 10*k* PMBC scRNA-seq Seurat object was downloaded from the link available in their v3.1 ATAC-seq Integration Vignette (https://satijalab.org/seurat/v3.1/atacseq_integration_vignette.html).

Cell labels from the scRNA-seq data were transferred using TransferData function based on the gene activity score. All the analyses were carried out using standard parameters. Jaccard similarities were evaluated using the scclusteval (v0.1.1) package[19].

Results

Limitations of *kallisto*-based analysis

At time of writing, *kallisto* does not natively support scATAC-seq analysis, though it can be applied to any scRNA-seq technology which supports CB and UMI. According to the *kallisto* manual, the technology needs to be specified with a tuple of indices indicating the read number, the start position and the end position of the CB, the UMI and the sequence respectively. In this sense, the technology specifier for standard 10x scRNA-seq with v2 chemistry is $0,0,16:0,16,26:1,0,0$ (see *kallisto* manual for details). Using this logic, a single fastq file contains sequence information and UMI is always required. scATAC-seq from 10x genomics is typically sequenced in paired-end mode and, moreover, there is no definition of UMI as reads can be deduplicated after genome alignment.

kallisto requires an index of predefined sequences, typically transcripts, to perform pseudoalignment and, if applied to scATAC-seq analysis, does not allow for any typical analysis in the epigenomic protocols, including the identification and quantification of enriched regions. Therefore, we computed an index on the genomic sequences for the 80,234 peaks identified by *cellranger-atac* and distributed together with fastq files. This ensures that the subsequent analysis were performed on the same regions and allowed us to quantify the bias, if any, introduced by *kallisto*.

kallisto primary analysis

We tested different strategies to overcome the technical limits and the absence of UMI. We evaluated concordance of different approaches in terms of adjusted mutual information (MI) of cell groups identified with a fixed set of filtering and processing parameters. Analysis based on *cellranger-atac* results is considered as ground truth. Results are reported in Table 1.

We tested two main strategies: in the first the R1 is pseudoaligned and the initial nucleotides of R2, cut at different thresholds, are used as UMI (pseudoUMI hereafter). As UMI is needed for deduplication, we reasoned that a duplicate in scATAC-seq should be identified by the same nucleotides, especially in the first portion of the read, where quality is higher. We observe generally high values of MI, with the notable exception of pseudoUMI 5*nt* long. Since basecall qualities are generally higher for R1 and *kallisto* does not use qualities in pseudoalignment, we also tested the strategy in which R2 is used for pseudoalignment and R1 is used to obtain pseudoUMI. Also in this scenario, 5*nt* pseudoUMI raised the worst results, while MI values were slightly higher than the forward configuration. In particular, we noticed the highest MI values when R2 is used

Table 1. Comparision of *cellranger-atac* and *kallisto* analysis. The table reports adjusted Mutual Information between single cell cluster assignments on *cellranger-atac* data and *kallisto* analysis. Different strategies to evaluate pseudoUMI are reported. All simulations raised high MI values, both in the forward and reverse approach, except for the pseudoUMI of length 5. The 10-Reverse configuration reached the highest score.

Comparison	Forward	Reverse
10x vs 5nt	0.1854	0.1733
10x vs 10nt	0.7434	0.7625
10x vs 15nt	0.7571	0.7398
10x vs 20nt	0.7356	0.7520
10x vs Random	0.7272	*None*

and pseudoUMI is 10*nt* long (*MI* = 0.7625). We also tested a configuration using R1 as sequence and 10*nt* UMI randomly generated. Interestingly, concordance remains in line with previous experiments (*MI* = 0.7272).

These data indicate that *kallisto* is able to properly quantify enrichments in scATAC-seq and does not introduce a considerable bias.

Analysis of DHS as reference

As one major limitation of a *kallisto*-based approach to scATAC-seq is the lack of peak calling routines and the need of a index of sequences for pseudoalignments, we reasoned that we could use any collection of regions that putatively would be target of ATAC-seq experiments. Since ATAC-seq is largely overlapping DHS we exploited the regions defined in the ENCODE project[20]. The DHS data provided by ENCODE includes 2, 888, 417 sites. We generated an additional dataset by merging regions closer than 500*bp* into 1, 040, 226 sites. We performed pseudoalignment on the full dataset, on the merged dataset and, lastly, on the full dataset summarized to the merged by *bustools* (see *Methods*). Pairwise comparison between performances of the three methods reveals lower values of MI (Figure 1B). Comparison with 10x data and the configuration 10-*rev* previously performed shows high values of MI when considering merged DHS intervals (*MI* = 0.7164 and 0.743 respectively). When pseudoalignmets are performed on the full DHS set, performance degrades to less than optimal levels. Since the number of DHS intervals is considerably higher than the typical number of regions identifiable by ATAC-seq, we tested the trend of MI at different cutoffs on the number of DHS included in the analysis (Figure 1C). MI reaches a plateau when approximately 50,000 regions are included into the analysis. This sets a reasonable target for region filtering during preprocessing stages of scATAC-seq data. In all, these findings support the suitability of using *kallisto* for identification of cell identities in scATAC-seq without any prior knowledge of the epigenetic status of single cells.

Identification of marker regions

A crucial step in the analysis of scATAC-seq data is the identification of marker peaks which can be used to functionally characterize different clusters. We tested the ability of our reference-based approach to identify differential DNase I hypersensitive sites that are overlapping or close to peaks identified with standard analysis. To this end, we first matched cell groups from DHS500 to groups identified after *cellranger-atac*. We selected the top 1, 000 peaks marking each DHS500 group and evaluated the concordance by mutual distance to the top 1, 000 significant markers in the matched groups ($p < 0.05$), we could identify significant markers only in five matched clusters. We found that the large majority of peaks (>= 80%) were overlapping between the two strategies or closer than 20*kb* (Figure 2). These results confirm the substantial equivalence between the standard strategy and the reference-based one.

Integration with scRNA-Seq data and cluster labeling

In addition to the analysis of technical suitability of *kallisto* for the analysis of scATAC-seq data, we investigated its validity in extracting biological insight. To this end, we performed a more detailed analysis of PBMC data by label transferring using Seurat V3[18], with the hypothesis that different approaches could lead to mislabeling of cells clusters. Matching is performed with the help of Gene Activity Scores calculated as sum of scATAC-seq counts over gene bodies extended 2*kb* upstream the TSS, Seurat's default approach. We applied the same transferring protocol on data derived from *cellranger-atac* counts and from the DHS500 approach (Figure 3), founding no relevant differences in the UMAP embeddings. A detailed quantification of cluster matches reveals a slight deviance in the characterization of NK subpopulations (Figure 4A). In addition to scores calculated by Seurat, we tested the ability of *bustools* summarization step to project and sum scATAC-seq values into Gene Activity using the identical mapping to extended gene bodies. In terms of cell labeling, this approach is equivalent to Seurat (Figure 4B), with the additional advantage of reduced run times.

Discussion/conclusions

Analysis of differential chromatin properties, through ATAC-seq and other quantitative approaches, relies on the identification of peaks or enriched regions, this is often achieved with the same statistical framework used in analysis of differential gene expression[21,22]. Identification of peaks is a key difference between the two approaches, *de novo* discovery of unannotated transcripts has been shown to be possible in early times of NGS[23], but the large majority of analysis is performed on gene models; conversely, analysis of epigenomes involves identification of regions of interest, although a large catalogues of such regions have been provided by several projects, such as the ENCODE project[24], the BluePrint project[25] or the GeneHancer database[26]. In single cell analysis, both scRNA-seq and scATAC-seq, identification of novel features may be an issue, especially because of the low coverage at which single cells are profiled. This work is the first, to our knowledge, to test the feasibility of a reference-based approach to ATAC-seq analysis,

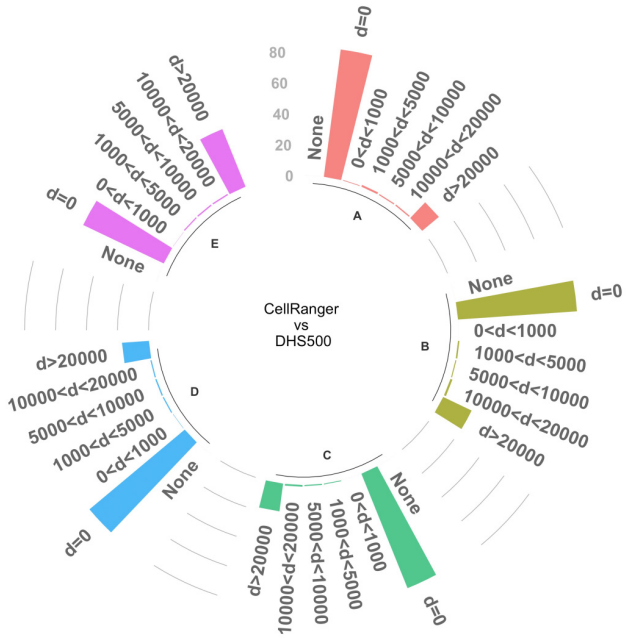

Figure 2. Analysis of peak concordance. The bars represent the proportion of marker peaks that are in common between DHS500 and *cellranger-atac*-based strategies at different distance thresholds. Only the top 1, 000 significant peaks ($p < 0.05$) were included in the analysis; the graph reports results for the 5 cell clusters (**A**–**E**) that contain the required amount of significant markers. The chart also reports the proportion of peaks without any match (*None*).

with a special focus on single cell ATAC-seq. In combination, we tested the suitability of *kallisto* to this kind of analysis, to maximise the performances of the whole process. Our results suggest that identification of cell groups using a reference-based approach is not different from a standard pipeline. Not only cells could be classified in a nearly identical way, but also differential features are largely matched between the analysis. The most obvious advantage is the gain in speed and efficiency: once reads have been demultiplexed, *kallisto* analysis requires short execution times, in the order of minutes, with limited hardware resources; this advantage has been known for a while and, in fact, it is has been demonstrated that it can be used on Rock64 hardware[27]. We also anticipate that adoption of a reference-based strategy comes with additional advantages, in particular functional annotations and gene associations are available for known regulatory regions[20] and, more recently, for DNase I Hypersensitive Sites[14]. Of course, our strategy has some limitations that come from the unavailability of read positioning on the genome: in addition to the impossibility of identifying novel peaks, it is not possible to perform some ATAC-specific analysis, such as nucleosome positioning or footprinting of transcription factors in accessible regions. Indeed, these two can be overcome if standard alignment is used in place of pseudoalignment, at cost of time. As concluding remark we would like to underline that, although we showed that *kallisto* can be effectively used for analysis of scATAC-seq data, we are aware that it has not been conceived for that purposes and, indeed, its interface needs some tweaks to make it work; for this reason we advocate the development of

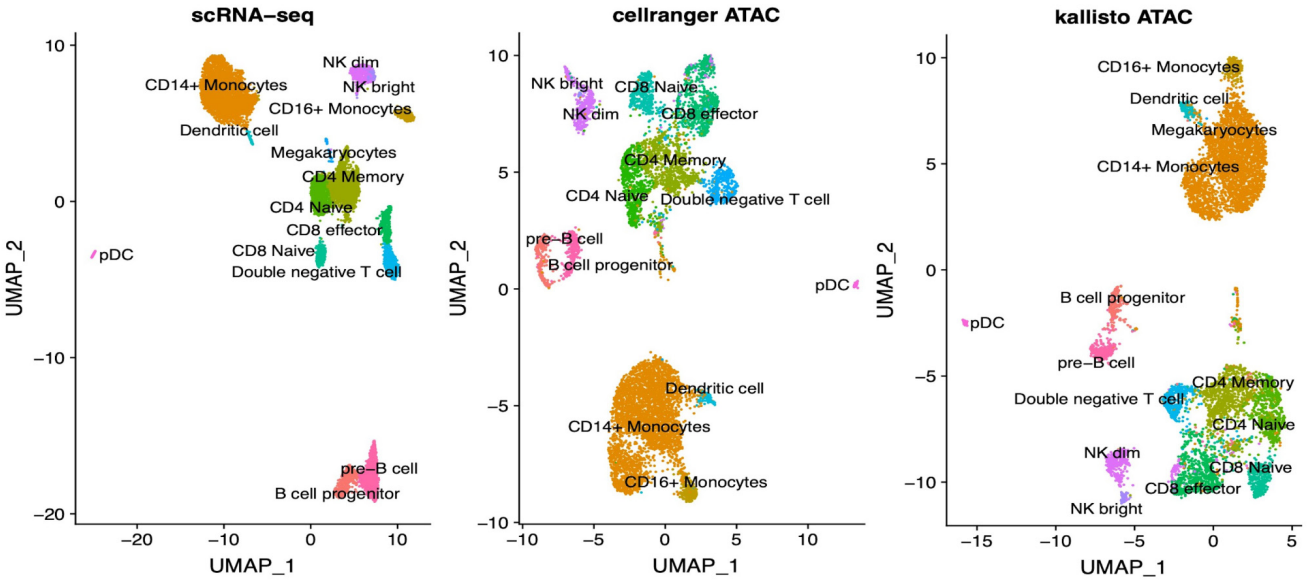

Figure 3. Results of label transfer from reference populations. The UMAP plot on the left represents scRNA-seq data of 10*k* PBMC as returned by Seurat vignette. The UMAP plots in the middle and on the right represent scATAC-seq analysis on *cellranger-atac* or *kallisto* analysis respectively. Cell clusters are consistent in their topology in the three plots, indicating the validity of *kallisto* for this kind of analysis.

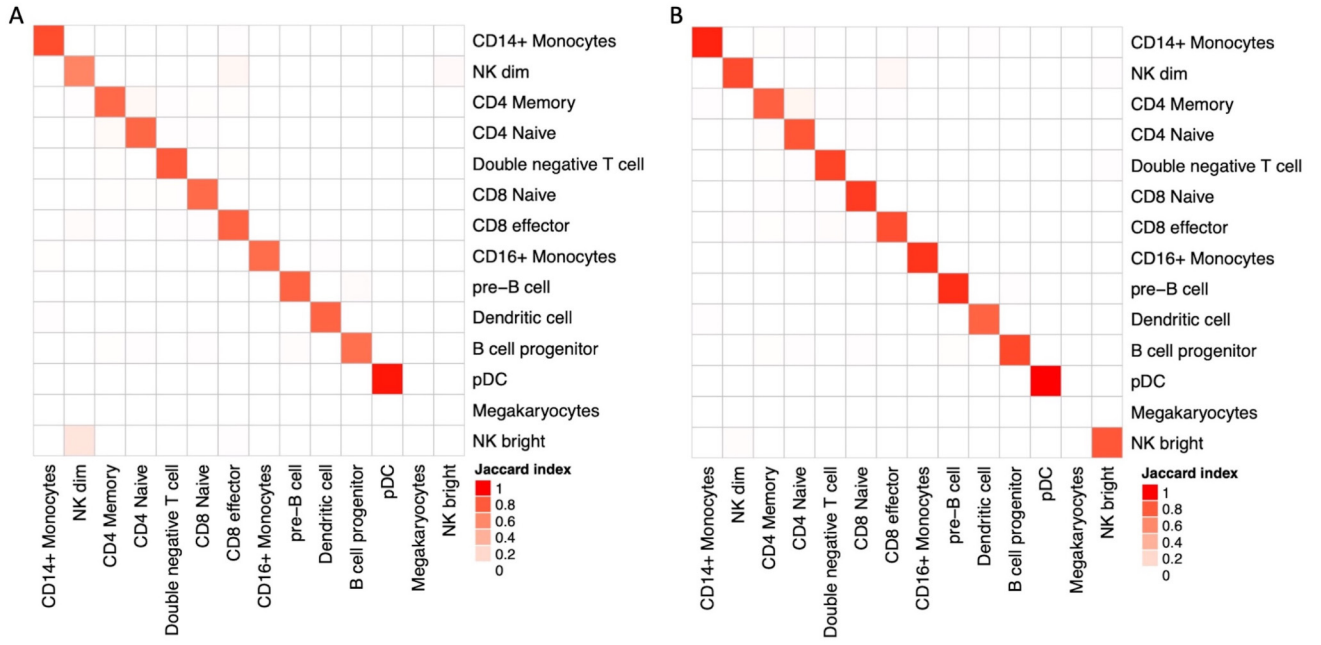

Figure 4. Analysis of Gene Activity Scores. (**A**) Pairwise Jaccard similarity between cell annotations as a result of label transfer from RNA-seq data using Gene Activity Score evaluated by Seurat. Concordance between results after *cellranger-atac* (rows) and DHS500 (columns) are largely comparable, with the notable exception of NK subpopulations. (**B**) Pairwise Jaccard similarity between cell annotations on DHS500 when Gene Activity Score is computed by Seurat (rows) or by *bustools* summarization step (columns).

tools which support scATAC-seq natively and other tools for postprocessing and data visualization.

Data availability

Source data

Single cell ATAC-seq data were downloaded from the 10x Genomics public datasets (https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_pbmc_10k) and include sequences for 10k PBMCs from a healthy donor. Access to the data is free but requires registration.

Extended data

Zenodo: vgiansanti/Kallisto-scATAC v1.0. https://doi.org/10.5281/zenodo.3703174[28].

This project contains a detailed explanation of the procedures described in this work and the list of DHS sites; this is also available at https://github.com/vgiansanti/Kallisto-scATAC.

Acknowledgements

The authors want to thank the people and supervisors who supported their work, in particular Giovanni Tonon, Catherine Dulac and Tim Sackton.

References

1. Svensson V, Vento-Tormo R, Teichmann SA: **Exponential scaling of single-cell RNA-seq in the past decade.** *Nat Protoc.* 2018; **13**(4): 599–604.
PubMed Abstract | Publisher Full Text
2. Wolf FA, Angerer P, Theis FJ: **SCANPY: large-scale single-cell gene expression data analysis.** *Genome Biol.* 2018; **19**(1): 15.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Zielezinski A, Vinga S, Almeida J, *et al.*: **Alignment-free sequence comparison: benefits, applications, and tools.** *Genome Biol.* 2017; **18**(1): 186.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Van den Berge K, Hembach KM, Soneson C, *et al.*: **RNA sequencing data: hitchhiker's guide to expression analysis.** *Annu Rev Biomed Data Sci.* 2019; **2**(1): 139–173.
Publisher Full Text
6. Conesa A, Madrigal P, Tarazona S, *et al.*: **A survey of best practices for RNA-seq data analysis.** *Genome Biol.* 2016; **17**(1): 13.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Harrow J, Frankish A, Gonzalez JM, *et al.*: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res.* 2012; **22**(9): 1760–1774.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Bray NL, Pimentel H, Melsted P, *et al.*: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotechnol.* 2016; **34**(5): 525–527.
PubMed Abstract | Publisher Full Text
9. Patro R, Duggal G, Love MI, *et al.*: **Salmon provides fast and bias-aware quantification of transcript expression.** *Nat Methods.* 2017; **14**(4): 417–419.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Melsted P, Ntranos V, Pachter L: **The barcode, UMI, set format and BUStools.** *Bioinformatics.* 2019; **35**(21): 4472–4473.
PubMed Abstract | Publisher Full Text

11. Zhang Y, Liu T, Meyer CA, *et al.*: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol.* 2008; **9**(9): R137.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Buenrostro JD, Giresi PG, Zaba LC, *et al.*: **Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.** *Nat Methods.* 2013; **10**(12): 1213–1218.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Thurman RE, Rynes E, Humbert R, *et al.*: **The accessible chromatin landscape of the human genome.** *Nature.* 2012; **489**(7414): 75–82.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Meuleman W, Muratov A, Rynes E, *et al.*: **Index and biological spectrum of accessible dna elements in the human genome.** *bioRxiv.* 2019.
Publisher Full Text

15. Sheffield NC, Thurman RE, Song L, *et al.*: **Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions.** *Genome Res.* 2013; **23**(5): 777–788.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis.** *Curr Protoc Bioinformatics.* 2014; **47**: 11.12.1–34.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Traag VA, Waltman L, van Eck NJ: **From Louvain to Leiden: guaranteeing well-connected communities.** *Sci Rep.* 2019; **9**(1): 5233.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Stuart T, Butler A, Hoffman P, *et al.*: **Comprehensive Integration of Single-Cell Data.** *Cell.* 2019; **177**(7): 1888–1902.e21.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Tang M: **crazyhottommy/scclusteval: second release for citing.** *Zenodo.* 2020.
Publisher Full Text

20. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, *et al.*: **Integrative analysis of 111 reference human epigenomes.** *Nature.* 2015; **518**(7539): 317–330.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol.* 2010; **11**(10): R106.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Yan F, Powell DR, Curtis DJ, *et al.*: **From reads to insight: a hitchhiker's guide to ATAC-seq data analysis.** *Genome Biol.* 2020; **21**(1): 22.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Robertson G, Schein J, Chiu R, *et al.*: ***De novo* assembly and analysis of RNA-seq data.** *Nat Methods.* 2010; **7**(11): 909–912.
PubMed Abstract | Publisher Full Text

24. ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature.* 2012; **489**(7414): 57–74.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Adams D, Altucci L, Antonarakis SE, *et al.*: **BLUEPRINT to decode the epigenetic signature written in blood.** *Nat Biotechnol.* 2012; **30**(3): 224–226.
PubMed Abstract | Publisher Full Text

26. Fishilevich S, Nudel R, Rappaport N, *et al.*: **GeneHancer: genome-wide integration of enhancers and target genes in GeneCards.** *Database (Oxford).* 2017; **2017**.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Tan QW, Mutwil M: **Inferring biosynthetic and gene regulatory networks from *Artemisia annua* RNA sequencing data on a credit card-sized ARM computer.** *Biochim Biophys Acta Gene Regul Mech.* 2019; 194429.
PubMed Abstract | Publisher Full Text

28. Giansanti V, Cittaro D: **vgiansanti/kallisto-scatac v1.0.** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.3703174