METHOD ARTICLE

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

Valentina Giansanti¹,², Ming Tang³, Davide Cittaro²

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

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
**Introduction**

Recent technological advances in single-cell technologies resulted in a tremendous increase in the throughput in a relatively short span of time. 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, 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.

More recently, analysis of NGS data benefit from technologies based on k-mer processing, allowing alignment-free sequence comparison. 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 to rely on a well-defined gene model such as GENCODE 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 or Salmon 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 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, opportunely tuned. Since ATAC-seq signal mirrors DNA accessibility as mapped by DNase-seq assays and catalogs of DNase I Hypersensitive Sites (DHS) are available, 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 10k PBMC from a healthy donor. We used the Peak by cell matrix HD5 (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, intervals closer than 500bp were clustered using bedtools.

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 (dHSfull) and an index for the merged set (dHS500). 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 16bp 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 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 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).

**Analysis of single-cell data**

Counts matrices were analysed using Scanpy (v1.4.2) 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. 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. Gene activity score was calculated using the CreateGeneActivityMatrix function or directly summarized.
The annotated 10k 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.htm).

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.

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 5nt 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, 5nt 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. Comparison 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 10nt long (MI = 0.7625). We also tested a configuration using R1 as sequence and 10nt 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\(^{12}\). The DHS data provided by ENCODE includes 2,888,417 sites. We generated an additional dataset by merging regions closer than 500bp 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 pseudoinmigments 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 20kb (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\(^{3}\), 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 2kb 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\(^{11,12}\). 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\(^{13}\), 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\(^{14}\), the BluePrint project\(^{15}\) or the GeneHancer database\(^{16}\). 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,
with a special focus on single cell ATAC-seq. In combination, we tested the suitability of \textit{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, \textit{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\textsuperscript{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\textsuperscript{10} and, more recently, for DNase I Hypersensitive Sites\textsuperscript{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 \textit{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...
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.

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.

Extended data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

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. Zieczinski 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 | Free 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

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).
Open Peer Review

Current Peer Review Status: 🟢 ❓

Version 1

Reviewer Report 06 May 2020

https://doi.org/10.5256/f1000research.25099.r62150

© 2020 Zhang Q. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Qiangfeng Cliff Zhang
MOE Key Laboratory of Bioinformatics, Beijing Advanced Innovation Center for Structural Biology, Center for Synthetic and Systems Biology, School of Life Sciences, Tsinghua University, Beijing, China

Comments:
The work by Giansanti et al. presents a novel and smart idea for scATAC-seq data analysis. It demonstrates the possibility of using a reference-based, pseudo-alignment method to reduce the computational requirement for scATAC-seq data analysis, with only a little sacrifice on precision. The idea is inspired by the using of pseudoalignment for bulk and single-cell RNA-seq quantification. Here they showed that with some tweaking of the input sequencing reads, they could use kallisto to analyze scATAC-seq data on a pre-defined set of DNase hypersensitive sites. They compared their results with the standard protocol (e.g. cellranger-atac) for peak quantification, single cell clustering, marker peaks identification, and gene activity score calculation.

The results very nicely revealed the consistency on peak quantification between kallisto-based method and cellranger-atac. The cell clusterings were almost identical between the new reference-based method and canonical mapping strategy. And the gene activity scores by two different methods also agreed well with each other. The approach presented in this study thus could be a very efficient way for scATAC-seq data analysis.

The following are a few comments/questions:
1. The method was only tested with one dataset - PMBC. In fact, single cell ATAC-seq data is usually very sparse. The PMBC dataset used in this study is of relatively high quality. The method remains to be tested on more datasets, especially on those of more sparse, lower-quality.

2. The key advantage of the method is presumably the much improved computational efficiency – there may be other advantages brought by the reference-based method. However, there is no results/statistics on the running time and memory usage in the manuscript. From the description, the improvement should be dramatic. I think it would be very nice to include a table or a figure to demonstrate the increase of computational efficiency. This could be a very helpful way to convince potential users.
3. As in the above, this whole strategy is so different. It is thus possible for the method to be used for some other scATAC-seq data analysis with advantages not only in computational efficiency. It would be good for the authors to explore.

4. The manuscript is well organized with the core ideas clearly described. But the presentation could be improved - there are a lot of very long sentences unnecessarily connected by “and”, “while”, etc.

5. The legend for Fig 1A says “The first (DHS) generated by kallisto on ~2M DNase I sites … ”, but according to the figure and the main text, it should be “~3M”?

Is the rationale for developing the new method (or application) clearly explained?  
Yes

Is the description of the method technically sound?  
Yes

Are sufficient details provided to allow replication of the method development and its use by others?  
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?  
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?  
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Bioinformatics, genomics, RNA structure, Genome structure, AI algorithms in biomedicine.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

**Author Response 20 May 2020**

**Davide Cittaro**, IRCCS San Raffaele Institute, Milan, Italy

Thank you for reviewing our manuscript and for the helpful comments. We have addressed major and minor points as detailed in the point-by-point response:

*The method was only tested with one dataset - PMBC. In fact, single cell ATAC-seq data is usually very sparse. The PMBC dataset used in this study is of relatively high quality. The method remains to be tested on more datasets, especially on those of more sparse, lower-quality.*
We agree that the PBMC dataset is of high quality. It was used as it could be considered a *de facto* standard in single cell analysis as it includes several populations at different degrees of separations (i.e. B-cells and T-Cells are well separated, while NK and CD8 are less clearly distinguished). We also would like to point out that it is difficult to identify low quality scATAC-seq datasets for two reasons: one is the relative novelty of this technique and the other is the positive bias in publications, which generally lack of negative or low-quality results. Nevertheless, we tried to address this question analyzing data for K562 cell line. Cell lines are supposedly more homogeneous, data were obtained on a low-throughput platform (Fluidigm C1). We believe that it could be considered a good example of “lower quality” dataset, compared to the PBMC, at least considering the information content. We show that our strategy is consistent with standard approaches based on alignment and peak identification, we can identify the same level of residual heterogeneity.

The key advantage of the method is presumably the much improved computational efficiency – there may be other advantages brought by the reference-based method. However, there is no results/statistics on the running time and memory usage in the manuscript. From the description, the improvement should be dramatic. I think it would be very nice to include a table or a figure to demonstrate the increase of computational efficiency. This could be a very helpful way to convince potential users.

Thank you for this comment. We benchmarked *kallisto+bustools* and compared it to *cellranger-atac*, the default application for 10x data. We added a dedicated section in the main text, which shows the large reduction in required resources. Note that the *cellranger-atac* pipeline includes several steps that are common in downstream analysis (such as Seurat or Scanpy). In order to make it fair, as explained in the text, we did not consider these steps in the comparison. In addition, we added runtime analysis for the approaches used in the analysis of K562 data.

As in the above, this whole strategy is so different. It is thus possible for the method to be used for some other scATAC-seq data analysis with advantages not only in computational efficiency. It would be good for the authors to explore.

Our work has been mainly motivated by the reduced resources that are needed by a *kallisto*-based approach, as we predict the number of scATAC-seq experiments will increase as well as the number of cells profiled. We anticipated additional advantages of a reference-based strategy in the first version of our manuscript, e.g. the availability of promoter-enhancer/gene interactions which could be readily applied to scATAC-seq data. During the revision process we had the opportunity to perform the analysis with non-optimal conditions (i.e. peak identification from single end reads instead of paired end), which led to slightly different results. This serendipitous finding suggests that our strategy, not relying on *de novo* identification, improves the stability of cell characterization and, therefore, the reproducibility of results. We added these observations in the discussion.

Of course, the usage of standardized reference could pave the way to a new class of processing steps not currently performed. As an example, one could identify a set of regions known to be generally accessible (or not) to perform standardized QC. Another example could be the identification of regions that could be used to score the cell cycle phases in scATAC-seq data, much like what is normally done with scRNA-seq data. We feel that all these examples require a deeper analysis, which is beyond the scope of this work, and any undemonstrated procedure would be, at best, greatly speculative. Our aim was to show general consistence between diverse approaches, which we believe has been demonstrated.
The manuscript is well organized with the core ideas clearly described. But the presentation could be improved - there are a lot of very long sentences unnecessarily connected by “and”, “while”, etc.

Thank you for this comment, we modified the text to increase readability.

The legend for Fig 1A says “The first (DHS) generated by kallisto on ~2M DNase I sites … ”, but according to the figure and the main text, it should be “~3M”?

Thank you for spotting the typo in the figure legend. We corrected accordingly.

**Competing Interests:** Nothing to disclose

---

Reviewer Report 30 March 2020

https://doi.org/10.5256/f1000research.25099.r61566

© 2020 Barozzi I. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Iros Barozzi**

Department of Surgery and Cancer, Imperial College London, London, UK

In their paper “Fast analysis of scATAC-seq data using a predefined set of genomic regions” Giansanti et al. suggest an efficient strategy to analyse scATAC-seq data using kallisto and bustools.

The paper is clearly written, the proposed strategy is well conceived and tested, and it will be useful for many researchers in the field of regulatory genomics. Clear advantages of this strategy are the reduced requirements in terms of computational resources and shorter execution times, when compared to other pipelines such as cellranger-atac. This comes at a cost, most notably the chance of missing signals at regions that are not present in the reference set. Nevertheless, in my opinion evaluations about this being a limitation has to be made on a case-by-case basis, and the authors clearly pointed this out (among other limitations) in the discussion. The authors also provide access to the full code, datasets and documentation to reproduce the analyses.

A wide range of parameters was tested, both in terms of handling and modifying the input sequences to make them suitable for kallisto, and in terms of pre- vs post- processing the genomic partition considered for indexing. Combinations that return results that are highly concordant with those obtained with cellranger-atac were highlighted. The authors then demonstrated the robustness of the biological inferences made using their strategy by showing a very large overlap with the results achieved by cellranger-atac (in terms of different groups of regions marking distinct clusters and clusters annotation based on label transferring from scRNA-seq data).

I am wondering if a natural application of this strategy would simplify the characterization of chromatin state at highly repetitive regions of mammalian genomes (e.g. indexing a database of transposable elements). This task would otherwise be quite difficult to handle explicitly with pipelines such as
cellranger-atac.

I only have two minor comments:

- Can the authors provide more details about the analysis described in the paragraph “Identification of marker regions”? How were the cell groups defined? How were the top 1,000 peaks for each group selected/identified?

- Fig. 1C: description of the blue curve seems to be missing.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Genomics; Transcriptional Regulation; Single-cell Transcriptomics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

---

**Author Response 20 May 2020**

**Davide Cittaro, IRCCS San Raffaele Institute, Milan, Italy**

Thank you for reviewing our work and for the comments. We have addressed your minor concerns as follows.

*Can the authors provide more details about the analysis described in the paragraph “Identification of marker regions”? How were the cell groups defined? How were the top 1,000 peaks for each group selected/identified?*

We apologize for lack of clarity in the manuscript. Cell groups were identified with the Leiden method, while markers were identified with Wilcoxon rank-sum test. The complete list of instructions used in the analysis is part of the repository linked in the main text, nevertheless we modified the text adding these specific details.
Fig. 1C: description of the blue curve seems to be missing.

Thank you for pointing this out. The blue line represented the fit of the DHS data. We acknowledge colouring scheme was not appropriate and, moreover, the fit DHS500 data was missing. In the revised manuscript we modified the figure accordingly.

**Competing Interests:** Nothing to disclose

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com