




# scATAC-seq preprocessing and imputation evaluation system for visualization, clustering and digital footprinting

Pavel Akhtyamov , Layal Shaheen, Mikhail Raevskiy, Alexey Stupnikov  and Yulia A. Medvedeva 

Corresponding author: Pavel Akhtyamov, Department of Biomedical Physics, Moscow Institute of Physics and Technology (National Research University), Institutskiy per., 9, 141701, Moscow Region, Russia; The National Medical Research Center for Endocrinology, Dm. Ulyanova, 11, 117036, Moscow, Russia.  
E-mail: akhtyamovpavel@gmail.com

## Abstract

Single-cell ATAC-seq (scATAC-seq) is a recently developed approach that provides means to investigate open chromatin at single cell level, to assess epigenetic regulation and transcription factors binding landscapes. The sparsity of the scATAC-seq data calls for imputation. Similarly, preprocessing (filtering) may be required to reduce computational load due to the large number of open regions. However, optimal strategies for both imputation and preprocessing have not been yet evaluated together. We present SAPIEnS (scATAC-seq Preprocessing and Imputation Evaluation System), a benchmark for scATAC-seq imputation frameworks, a combination of state-of-the-art imputation methods with commonly used preprocessing techniques. We assess different types of scATAC-seq analysis, i.e. clustering, visualization and digital genomic footprinting, and attain optimal preprocessing-imputation strategies. We discuss the benefits of the imputation framework depending on the task and the number of the dataset features (peaks). We conclude that the preprocessing with the Boruta method is beneficial for the majority of tasks, while imputation is helpful mostly for small datasets. We also implement a SAPIEnS database with pre-computed transcription factor footprints based on imputed data with their activity scores in a specific cell type. SAPIEnS is published at: <https://github.com/lab-medvedeva/SAPIEnS>. SAPIEnS database is available at: <https://sapiensdb.com>

**Keywords:** single cell; scATAC-seq; imputation; preprocessing; digital footprinting

## INTRODUCTION

Epigenetic regulation and transcription factor (TF) binding represent the two critical components of transcription regulation machinery. Assay of Transposase Accessible Chromatin (ATAC-seq) [1] is a sequencing-based approach for the global discovery of open chromatin, a distinctive feature of active regulatory regions, including TF binding sites (TFBS). However, ATAC-seq does not allow for the identification of a specific TF or any other regulators. In order for the experimental methods (such as chromatin immunoprecipitation or ChIP-seq) to detect a specific TF, they require a high input number of cells [2]. In addition, ChIP-seq is limited to one TF per assay and is further restricted to those TFs, for which antibodies are available. Therefore, direct experimental detection methods remain costly, or even impossible, to

study the binding of multiple TFs in parallel. Digital genomic footprinting (DGF) [3, 4]—a computational approach to process chromatin accessibility assays such as DNase-seq [5] or ATAC-seq [6]—can overcome some of the limitations of ChIP-based methods. DGF is based on the observation that a TF being bound to DNA protects it from cleavage, resulting in local regions of decreased accessibility. ATAC-seq protocol could be scaled to a single-cell level allowing detection of rare cell populations and transition states. In combination with DGF, scATAC-seq allows the detection of TFBS at the level of a single cell.

One of the main problems of the single-cell data is sparsity: single-cell RNA-seq has 70–90% of zero counts and the problem of the scRNA-seq imputation data has been widely studied [7]. The problem escalates to a larger scale in case of scATAC-seq,

---

**Pavel Akhtyamov** is a PhD student at Moscow Institute of Physics and Technology, Russia; researcher of a Group of Medical bioinformatics and Omix technology at the National Research Center of Endocrinology, Russia. He is jointly supervised by Yulia Medvedeva and Alexey Stupnikov. He is currently focused on developing pipelines of processing and evaluation for single-cell omics datasets.

**Layal Shaheen** is a Mathematical Biology and Bioinformatics PhD student at Moscow Institute of Physics and Technology, Russia. Her work is devoted to prediction of multifactorial disease using polygenic risk scores.

**Mikhail Raevskiy** is a PhD student at École Polytechnique Fédérale de Lausanne, Switzerland. His work is mainly related to assessing new single-cell imputation and pseudo-time analysis methods.

**Alexey Stupnikov** is postdoctoral researcher at Moscow Institute of Physics and Technology; researcher of a Group of Medical bioinformatics and Omix technology at the National Research Center of Endocrinology, Russia. He is currently focused on development of new single-cell evaluation metrics.

**Yulia Medvedeva** is a head of the Group of Regulatory Transcriptomics and Epigenomics at the Research Center of Biotechnology, Russian Academy of Sciences; of a Group of Medical bioinformatics and Omix technology at the National Research Center of Endocrinology; and of the Laboratory of Bioinformatics of Cell Technologies at Moscow Institute of Physics and Technology, Russia. Her work is mainly focused on the regulatory genomics and transcriptomics.

**Received:** August 24, 2023. **Revised:** October 29, 2023. **Accepted:** November 14, 2023

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

which typically produces only 3–7% of non-zero values in peak-cell matrix [8]. However, scATAC-seq data are a valuable source of information on chromatin regulation and it can in turn be used for the scRNA-seq data imputation [9]. Thus, scATAC-seq imputation methods are critical for the downstream analysis.

The term ‘imputation’ may refer to the whole procedure of data transformation for sparsity reduction or can be attributed only to the key step of this process [10]. In this study, we apply the term ‘imputation framework’ the complete procedure of counts transformation and refer to the second step of this procedure as ‘imputation’. Generally, an imputation framework for scATAC-seq data involves preprocessing, imputation and postprocessing.

Due to a typically very large number of initially identified peaks in scATAC-seq, a preprocessing step is sought to reduce it in order to optimize future steps in terms of computational memory resources while limiting the inevitable loss of signal. In this study, we tested Boruta [11], a machine learning method for feature selection, and Cicero [12], a bioinformatic tool that detects co-accessible chromatin elements to identify relevant characteristic chromatin accessibility patterns that can serve as cell-type-specific markers.

Imputation, a key step of any imputation framework, is a mathematical model allowing for data transformation with sparsity reduction and dropout recovery. Up to date, three approaches are available for imputation of scATAC-seq data, e.g. SCALE (Single-Cell ATAC-seq analysis via Latent feature Extraction) [13], scOpen [8] and cisTopic [10]. scOpen is an unsupervised learning model for scATAC-seq data imputation. It estimates accessibility scores to indicate if a region is open in a particular cell based on a non-negative matrix factorization (NMF), which makes no assumption on the data distribution. SCALE combines a deep generative framework and a probabilistic Gaussian Mixture Model to learn latent features that accurately characterize scATAC-seq data. SCALE uses the latent features to cluster cell mixtures into subpopulations and to denoise and impute missing values in scATAC-seq data. SCALE requires a graphics processing unit (GPU) for training which limits the number of features (peaks) to be analyzed due to a typically small size of GPU memory. cisTopic is a Bayesian-based method reported to have an exponential increase in the running time for an increasing number of reads; therefore, we excluded it from the benchmarking. Postprocessing is implemented in some of the recent imputation frameworks [8, 13] to allow the selection of the optimal candidate transformation out of several produced in the second step in these frameworks.

Several comparison studies conducted direct [14] or indirect [8, 13] benchmarking, or cross-referencing scATAC-seq results with relevant scRNA-seq [15–17], driving conclusions on imputation step quality for scATAC-seq.

Still, the most recent papers [15, 16] employ imputation frameworks as a preprocessing step for general scATAC-seq pipeline evaluation (gene scoring [15], single-cell integration [16]) with no focus on the impact of the different imputation strategies. Independent benchmark studies [14, 17] and method presenting papers [8, 13] have provided effective benchmarking protocols for comparing scATAC-seq imputation methods; however, these works do not explicitly investigate the contribution of preprocessing and postprocessing steps, which can have a significant impact on the imputation framework results. Recent work from Liu and colleagues [17] benchmarked thoroughly the effects of imputation on scATAC-seq downstream analysis, but in this work, only application of scRNA-seq imputation methods was considered. Thus, the evaluation of imputation framework performance accounting for the preprocessing

and postprocessing methods using state-of-the-art scATAC-seq-specific imputation approaches remains an unaddressed problem.

In this study, we present scATAC-seq Preprocessing and Imputation Evaluation System (SAPIEnS), a novel benchmarking approach and instrument, to address missing points in previous benchmarks. We evaluated two recently introduced imputation methods: SCALE [13] and scOpen [8] in combination with several preprocessing approaches (a fixed threshold, Boruta [11] and Cicero [12]) and postprocessing strategies. To perform an evaluation of the imputation frameworks, we applied not only statistical metrics for clustering and visualization but also the quality of DGF in a well-studied and validated biological system of haematopoiesis. Results of the DGF based on the imputed data are provided in SAPIEnS database.

## METHODS

### Design of the benchmark

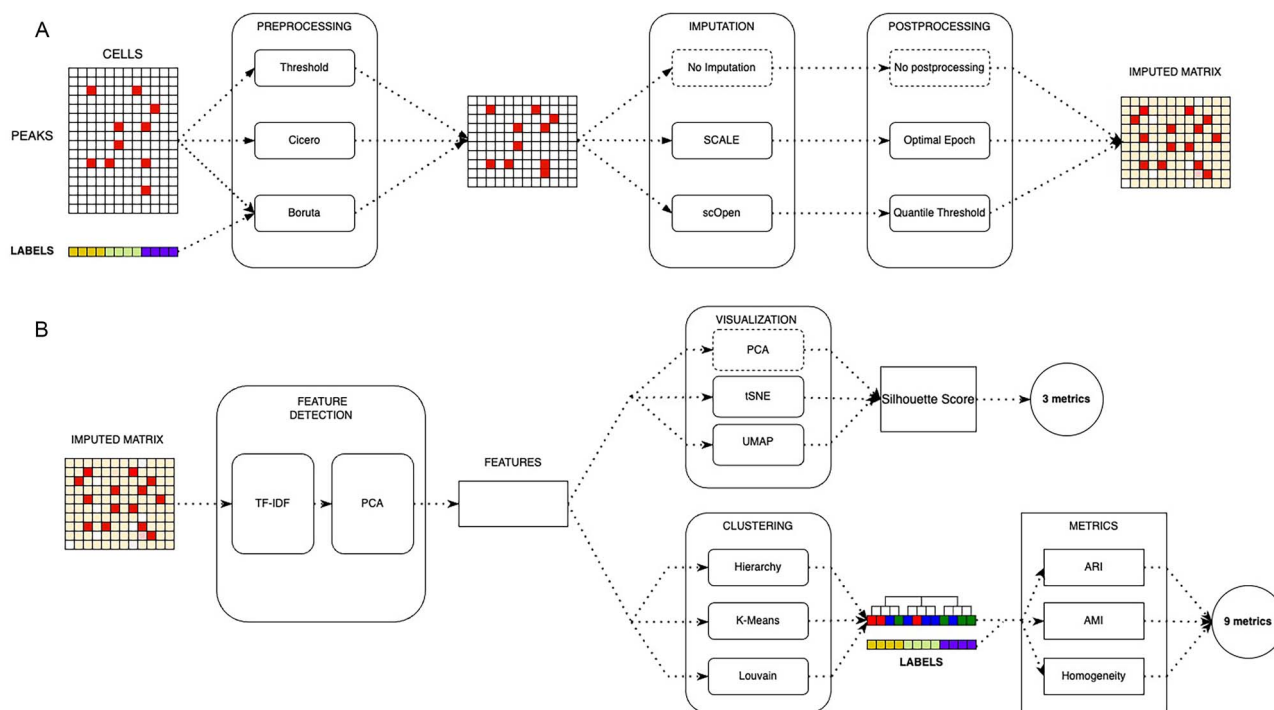
In brief, SAPIEnS consists of two components: an scATAC-seq imputation framework (Figure 1A) and an evaluation procedure (Figure 1B). In turn, the former component has three major parts: preprocessing, imputation and fine-tuning (Figure 1A). Preprocessing starts from the count matrices and selects peaks using one of the three methods: Threshold, Cicero [12] and Boruta [11]. These matrices can be directly used for the downstream analysis or can be subjected to the imputation methods: scOPEN [8] and SCALE [13]. Each imputation method was subjected to the corresponding postprocessing (fine-tuning of hyperparameters). Therefore, we obtained and benchmarked nine imputation frameworks (Figure 1A).

Various approaches may be employed to assess clustering results quality and properties that typically focus on clusters’ compactness and separation [18, 19], or discriminative power of the selected features [20, 21]. To balance different quality metrics, we designed the procedure for assessing the imputation quality as follows. We projected scATAC-seq imputed matrix using TF-IDF transformation (Term Frequency - Inverse Document Frequency) [22] and extracted 150 PCA components, following the protocol suggested in [23]. After that, we evaluated the features with the three most common methods for visualization (PCA, UMAP, tSNE), and three methods for clustering (Hierarchy, K-Means, Louvain) coupled with 4 common statistical metrics [Silhouette score for visualization and ARI (Adjusted Rand Index), AMI (Adjusted Mutual Info) and Homogeneity score for clustering] (Figure 1B). For the majority of metrics, no major difference between the original imputed matrix and the one after a PCA was observed (Supplementary Figure 1a-1). Moreover, in many cases, the original imputation matrix demonstrated slightly worse results (Supplementary Figure 1m,n), e.g. the Silhouette score reached saturation at about 50 principal components (see Supplementary Figure 1o), confirming 150 components to be adequate or even outperforming the complete set of features.

Next, we estimated the TFBS detection improvement after applying all imputation frameworks using DGF activity scores for a predefined validated set of transcription factors.

### Imputation frameworks

Preprocessing selects  $N$  peaks from data with the strongest signal. We used  $N = 50\,000$  peaks for datasets that have more than 100 000 peaks and  $N = 10\,000$  peaks for smaller datasets. The threshold approach selects peaks based on the number of cells where the peak has been detected. Cicero extracts peaks with the top co-accessibility score with other peaks in the dataset (executed



**Figure 1.** (A) Imputation frameworks general structure: one of the three preprocessing methods is followed by one of the three imputation approaches and the corresponding postprocessing steps. (B) Design of the procedure for the imputation framework quality assessment.

by each chromosome independently to reduce memory package). Boruta method chooses top peaks based on their impact on the detection of predefined labels. Annotation was obtained from FACS-sorting or labels transferred from scRNA-seq depending on the original dataset. We used `perc` option to relax the quantile thresholds and to select  $N$  relevant peaks.

The imputation methods were executed with the following parameters: Raw (no imputation is used), SCALE (`-latent 10 -min_peaks 1 -x 0 -encode_dim 1600 600 300 100 -max_iter 100000`, disabled early stopping procedure, dumped binarized imputation matrix every 10 000 iterations), Parameters `-x` and `-min_peaks` were set to disable preprocessing step in the SCALE package, `-encode_dim` was reduced in order to optimize for the GPU RAM for the large datasets, and the rest are the default set-up. scOpen was executed with default parameters except the `-binary_quantile` parameter, see below.

Imputation methods rely on hyperparameters for modeling peak representations (embeddings). For SCALE, we compute all 12 statistical metrics at each 10 000 up to 100 000 iterations. For each metric, we select the iteration with the best score. For scOpen method, we select a binary quantile threshold as a hyperparameter varying from 0.0 to 1.0 with step 0.1. The optimal hyperparameter is selected by majority voting procedure for the 12 winners in both methods (Supplementary Figure 3a).

## Data sampling

Data subsampling is most commonly employed to study such properties as robustness [24, 25] and reproducibility [26, 27]. We have performed subsampling of peak counts for the large datasets to confirm the observations made for the large and the small datasets. Thus, new datasets of a smaller size than the original one were simulated. Imputation frameworks could provide bias when selecting different numbers of peaks for small and large datasets. For three large datasets with different count order of

peaks (100 000 peaks, 230 000 peaks, 467 000 peaks), we performed sampling of 20, 40, 60 and 80% non-zero peaks of the original count matrix using a random function. For every subsampled matrix, we executed SAPIEnS and aggregated the results by every metric. Sampling dataset to 20% peaks simulates the behaviour of small datasets.

## Footprints-based imputation metrics

Clustering or visualization-based metrics do not reflect all the impact of the imputation frameworks. They also should lead to increased biologically interpretability, such as improved DGF output. We designed and applied strategy for estimating DGF quality based on activity scores of genomic footprints obtained by RGT-HINT [28] in a well-studied haematopoietic system [23] on pseudo-bulk data. To match footprints to TFBS, we used HOCOMOCO motifs [29].

For a set of key TFs linked to haematopoiesis, i.e. ELF1, FLI1, GATA-family, IKZF1, RUNX1, SPI1, TAL1, LYL, ERG and ETS (11 TFs in total), we calculated ranks of the activity scores for TF digital footprints in HSC to CLP lineage, i.e. between HSC (Haematopoietic Stem Cells), MPP (Multipotent progenitors), LMPP (Lympho-myeloid primed progenitor) and CLP (Common lymphoid progenitors) pseudo-bulk ATAC-seq (Supplementary Table 2).

The evaluation procedure was implemented as follows. For every cell type and every TF, we calculated an activity score with RGT-HINT tool [28]. Next, for every considered cell type transition and each TF, we computed the difference in activity scores and ranked these values. This resulted in a vector of ranks corresponding to the TFs' relative changes of activity scores. Afterwards, for every TF, we calculated a difference between the TF's rank in the case of imputation and the baseline (a threshold preprocessing with no specific imputation), resulting in a TF rank improvement score (TFRIS). Out of 11 TFs, GATA3 and IKZF1 are expected to be

**Table 1:** Datasets used in the benchmark

Dataset name	Reference	Number of cell types	Number of cells	Number of peaks	Reason to be included
BreastTumor	[30]	4	384	27 884	Batch effect
Forebrain	[31]	8	2088	11 286	Complex subclusters structure
HSC	[23]	11	2034	230 000	Used for footprinting analysis
FibroCard	[32]	9	79 514	287 000	Complex subclusters structure
MouseAtlas	[33]	30	80 000	467 000	Two levels of annotation
T Cells	[34]	4	765	8415	Clear cell differentiation
CellLines	[35]	6	1224	13 464	Used in scOpen and SCALE original papers
PBMC5K	[36]	10	5000	97 998	Labels were transferred from scRNA-seq

markers of more differentiated lymphoid cells, while other nine TFs are markers of progenitor cells. Therefore, one would expect nine markers of the progenitor cells to be highly ranked and two markers of the differentiated cells to be lowly ranked. To produce a comparable metric, we inverted the ranking for differentiated lymphoid cell markers, GATA3 and IKZF1. Next, we summed up rankings of all TFs with a threshold preprocessing and no imputation as a baseline in comparison with the results of other imputation frameworks, resulting in a method rank improvement score (MRIS) (Figure 4A).

The described pipeline allows to rank imputation frameworks. We split the 11 tested TFs into two groups depending on whether their rank increased or decreased after the imputation framework had been applied. For every method, we calculated the sum of the increase in ranks. The final score for an imputation framework is the sum of the ranks gained for a specific branch compared with the baseline method (Maximum MRIS).

## Data description

We benchmarked all the imputation frameworks on eight datasets (Table 1, Supplementary Materials). We deliberately selected four relatively small datasets (less than 100k peaks) and four large datasets (at least 100k peaks) to explore the impact of the dataset feature number on the imputation frameworks.

## Implementation of the pipeline

Pipeline wrappers were implemented using Python 3.7, Bash 4.4, R 4.1.3. Clustering metrics were obtained from ‘umap-learn’ package [37] and ‘scanpy’ [38] method ‘louvain’ [39] with ‘scikit-learn’ [40] integrated metrics ARI, AMI, Homogeneity Score. Every step of the imputation framework has been completed on SLURM-based cluster on computed nodes with 16 CPU and 128GB RAM. We used the borutaPy package for Boruta implementation [41]. scOpen (version 1.0.0) and SCALE (version 1.0.1) packages were obtained at github.com. The SAPIEnS code is available at <https://github.com/lab-medvedeva/SAPIEnS>.

## SAPIEnS database

SAPIEnS database has been implemented with Django Framework. Front-end was implemented in JavaScript with `table-filter.js` and `d3.js` for visualizing digital footprinting graph interactions.

## RESULTS

### Small datasets strongly benefit from imputation

We applied all imputation frameworks to eight datasets, four small and four large ones. Typically, datasets with a lower sequencing depth produce a smaller number of peaks due to a lower statistical power. We consider the threshold preprocessing with no imputation (column `threshold`) as a baseline for the

comparisons. Surprisingly, only small datasets benefit from imputation in terms of both visualization and clustering metrics (Figures 2 and 3A–C). Although Friedman test results show a significant difference between groups of metrics derived for SCALE, scOpen and no imputation for both large and small datasets, post-hoc pair-wise analysis shows imputation improvement for small datasets and no improvement for large datasets (Supplementary Table 1). Boruta preprocessing in combination with SCALE imputation improves visualization with UMAP (Figure 2A, Supplementary Figure 4), while in terms of the clustering metrics, both imputation methods perform rather well (Figure 2B and C). On the other hand, the preprocessing without imputation decreases the clustering quality probably due to suboptimal peak selection.

In more detail, all the label-based metrics (AMI, ARI, homogeneity) demonstrate increased quality of Hierarchical and K-Means clustering when an imputation framework is applied. However, for the Louvain clustering, the increase is not that pronounced. Although imputation frameworks with SCALE often infer the data structure finer, the increase over the results of imputation frameworks with scOpen is not that dramatic, suggesting that the choice of imputation method should not be the main priority for small datasets.

### Large datasets typically do not benefit from imputation but from preprocessing

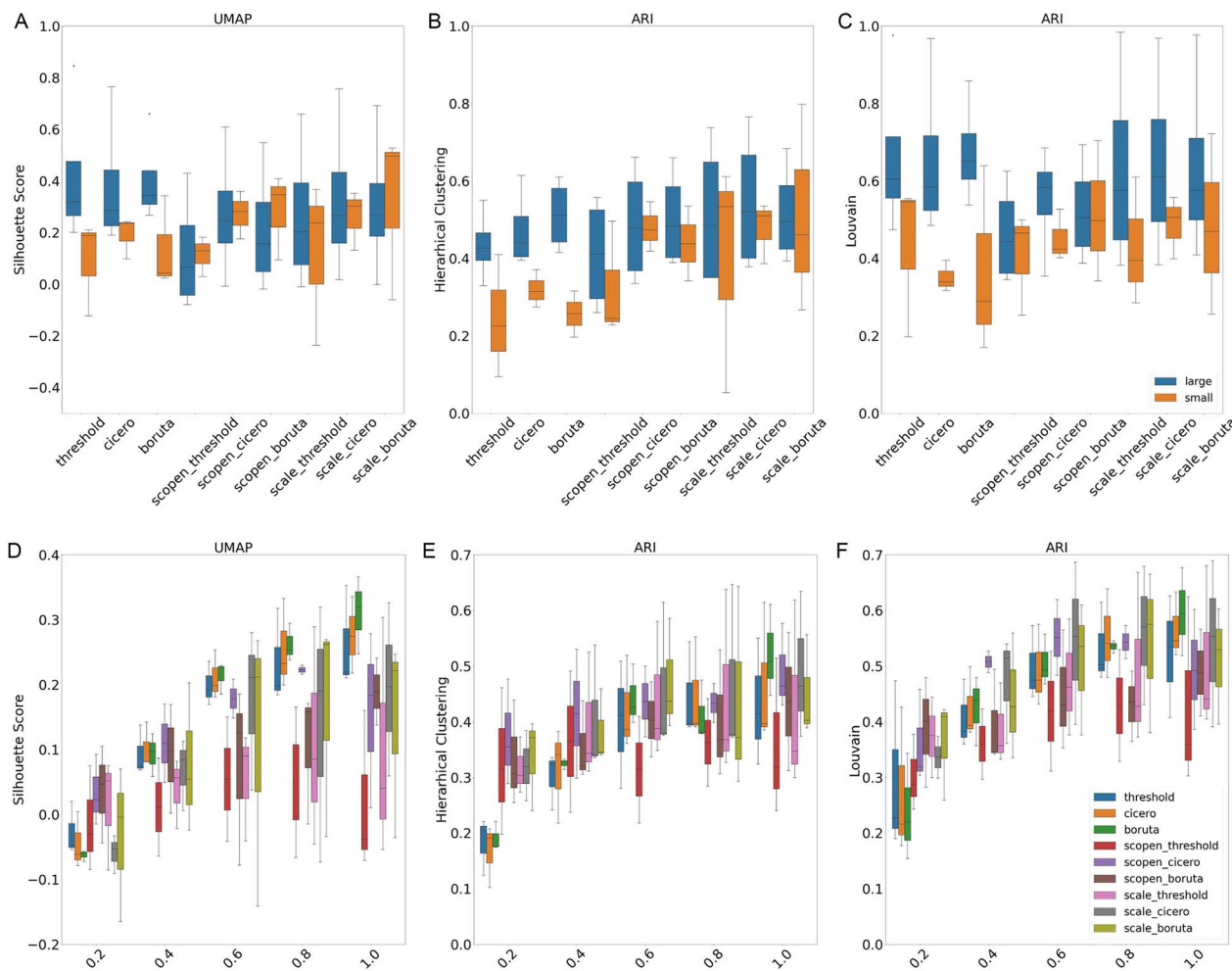
As expected, the majority of the metrics show that large datasets have an advantage over small datasets (Figure 2, Supplementary Figure 4). However, none of the imputation approaches improve quality metrics significantly. The metric for the Hierarchical (Figure 2B) and K-Means clustering (Supplementary Figure 4) remains almost the same. Moreover, the scores in the case of Louvain clustering dropped after the imputation with scOpen while remaining the same after the imputation with SCALE (Figure 2C). Visualization quality also drops after imputation (Figures 2A and 3D–F).

Additionally, we performed subsampling of 20, 40, 60 and 80% non-zero peaks of the original dataset to further validate the difference between small and large datasets. For deeper subsampling levels, metrics drop significantly holding the originally observed patterns: for small datasets (20% of the original), imputation improves both clustering and visualization metrics, while for bigger datasets (80–100%), imputation is not beneficial for either clustering or visualization.

### Footprinting of well-known haematopoietic regulators benefits from both preprocessing and imputation

Although clustering and visualization metrics provide an important quality estimate and give insight into the data structure, they may not validly reflect the optimal parameters for biologically





**Figure 2.** Benchmarking results splitted by the size of the datasets (large with more than 50 000 peaks and small with less than 50 000). (A) Silhouette scores of UMAP embeddings; (B) ARI scores for hierarchical clustering; (C) ARI scores for Louvain clustering; (D) Silhouette scores alterations with peaks subsampling; (E) ARI scores for hierarchical clustering alterations with peaks subsampling; (F) ARI scores for Louvain clustering alterations with peaks subsampling.

relevant downstream analysis. One of the key applications of the scATAC-seq data is the detection of active regulatory regions where transcription factors can bind. Digital footprinting provides an illustration of the role of the imputation framework on the quality of TFBS detection.

We selected a haematopoiesis dataset with an increased share of progenitor cells [23] and focused on HSC to CLP lineage with well-known regulators (see Methods). However, a comparison of HSC and CLP directly may not reflect all the changes in the lineage. Therefore, we included additional intermediate cell types of the development branch (MPP and LMPP) and obtained activity scores for digital footprinting in 4 transitions (HSC to MPP, MPP to LMPP and LMPP and CLP). The clustering of four lineage stages based on TF activity scores is in concordance with the similarity of the cell types (Supplementary Figure 5), suggesting that the set of the TFs is biologically reasonable.

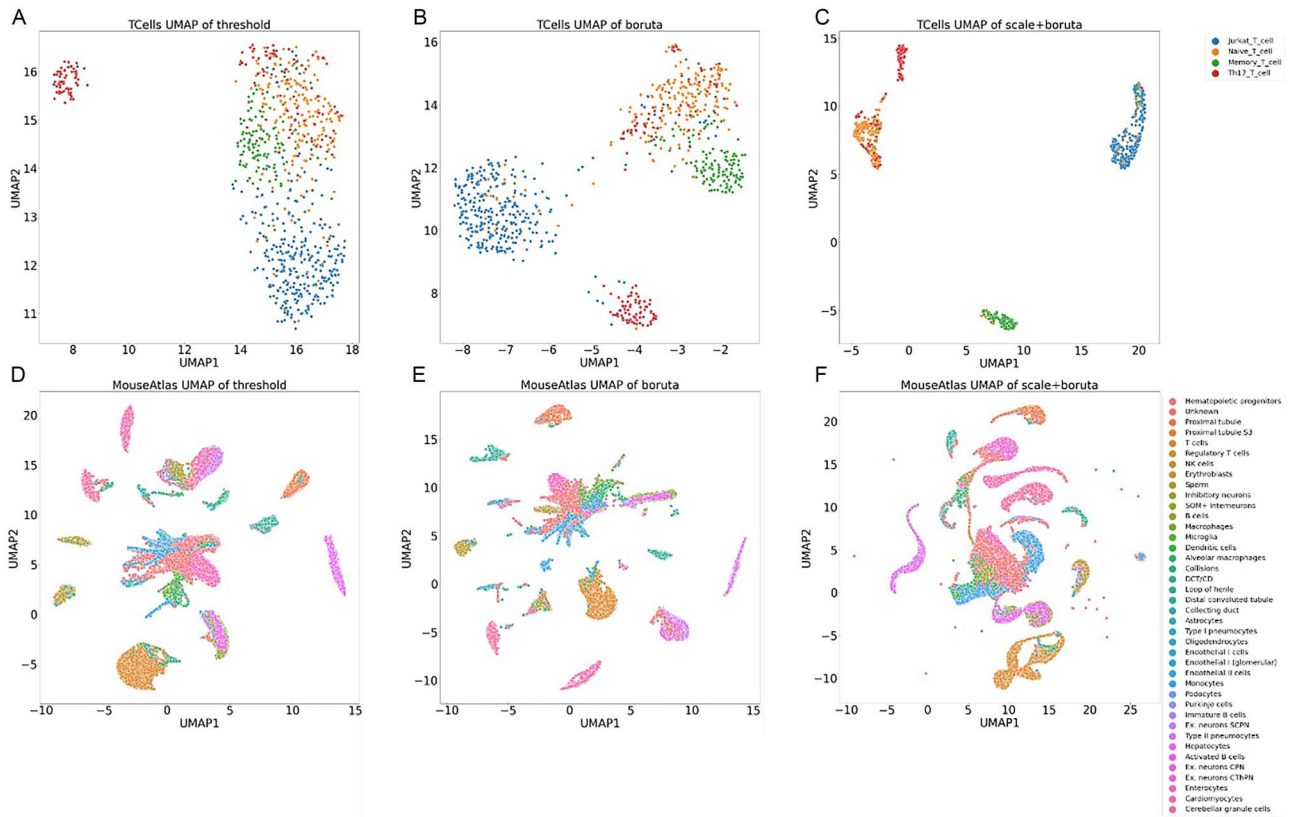
We detected three TF groups with the highest TFRIS: FLI1, GATA2, LYL1, ERG for HSC → MPP branch (Figure 4B), ETS1, GATA3, IKZF1, RUNX1, TAL1 for MPP → LMPP branch (Figure 4C) and ELF1, SPI1 for LMPP → CLP branch (Figure 4D). Finally, an aggregated score—maximum MRIS (see Methods)—reveals the improvement of imputation and preprocessing methods for specific transitions (Figure 4E).

First, we investigated footprinting profiles between two cell types with the highest rank improvements (see Methods) for separate TFs. For FLI1, the profile is increased in HSC [42] (Supplementary Figure 6a), while for IKZF1, the profile is increased in CLP [43] (Supplementary Figure 6b).

Second, we investigated the rank improvements for all TFs (Figure 4B–E). Boruta preprocessing with SCALE imputation provides the highest summarized activity score, while Boruta with scOpen gives the second-best results. On the other hand, scOpen imputation with threshold preprocessing reveals downgrade ranking (Figure 4F).

Third, we repeated the clustering of the cells based only on peak signal for the 11 selected TFs. The mean label score improvement (MLSI) was calculated as an improvement of average scores obtained from nine metrics (AMI, ARI, Homogeneity score with a combination of K-Means, Louvain and Hierarchical Clustering Methods) over the threshold baseline. The results support Boruta with SCALE as the most beneficial imputation framework (Figure 5A).

Fourth, to estimate the concordance between the quality of the footprinting and the quality of clustering, we correlated the maximum MRIS with the MLSI. Both Kendall- $\tau$  and Pearson correlation are relatively high suggesting the concordance of



**Figure 3.** Sample UMAP visualizations for a small Tcells (A–C) and a large MouseAtlas (D–F) datasets.

clustering and footprinting metrics (Figure 5B). Of note, Kendall- $\tau$  and Pearson are increased for the TF-limited clustering (Figure 5B and C).

Clustering on the TF-limited peaks emphasizes the success of the imputation frameworks with Boruta preprocessing (Figure 5C). However, even if Boruta preprocessing cannot be applied, Cicero is the second method of choice, while threshold-based methods should be avoided.

### SAPIEnS database simplifies search of transcription factors after imputation

To present the results of TFBS footprinting, we designed a SAPIEnS database (Supplementary Figure 7). The database contains the results of digital footprinting between all cell type pairs, obtained with all imputation frameworks to allow users to make their own choices based on biological interpretability.

SAPIEnS provides several modes of data analysis. First, it is possible to select two clusters in one experiment and compare the results of various imputation frameworks for a TF of the user's choice. This mode allows the user to verify whether the generally best-performing imputation framework (e.g. Boruta + SCALE) is indeed the best-performing framework for a TF of interest. Second, it is possible to select two clusters from an experiment with a fixed imputation framework and detect all TFs with significant footprints between two clusters. This mode may allow a user to determine novel TF regulators for a contrast of interest. The second mode of analysis can be expanded to cell lineage that consists of more than two clusters. Third, it is possible to select a TF in the experiment with a fixed imputation framework and find clusters of cells with high or low activity of these TFs. This mode may provide a hint in which cell types a TF of interest actually works. Finally, the user can get summary statistics of the

experiments: the number of significant TFs for each experiment grouped by the HOCOMOCO motif quality type (A, B, C and D). The database is available at <https://sapiensdb.com>.

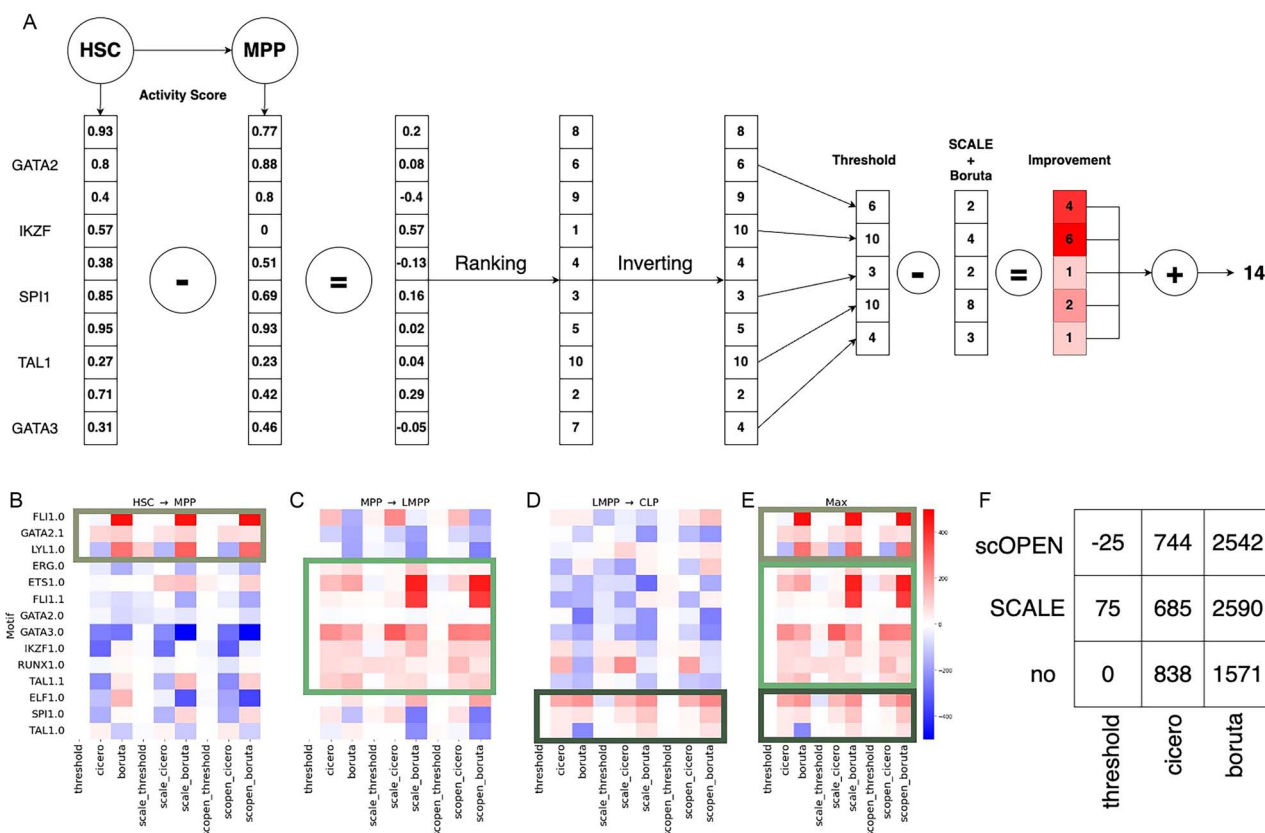
## DISCUSSION

scATAC-seq is a recently introduced approach that has been proved effective for the analysis of a cell regulatory landscape. Due to a large number of open chromatin regions detected in the scATAC-seq data, and a limited sequencing depth, imputation and pre-processing (filtering) are considered critical steps in scATAC-seq analysis. Several studies have evaluated imputation approaches and preprocessing techniques separately; however, assessment of a whole imputation framework, including both preprocessing methods and the imputation algorithms, has not yet been carried out.

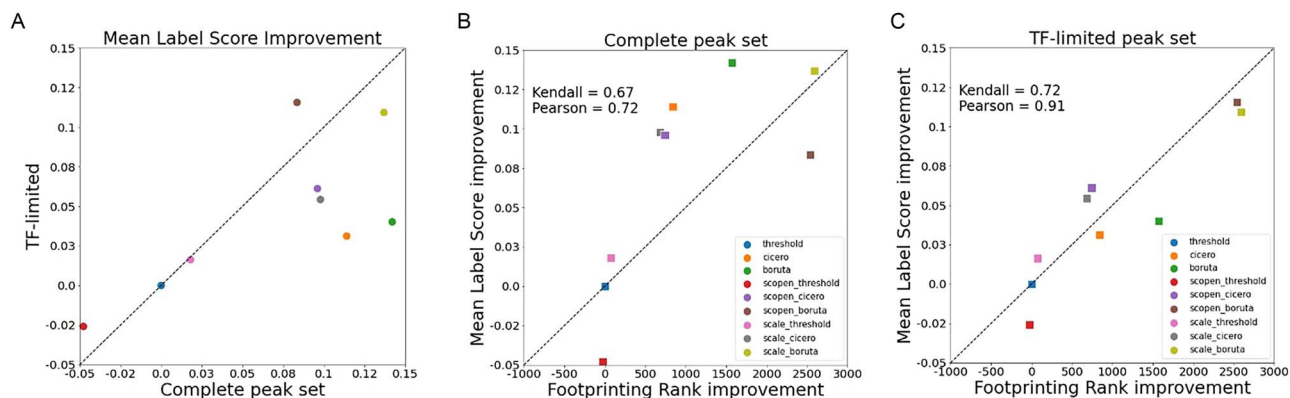
We introduce SAPIEnS, a comprehensive strategy for scATAC-seq imputation frameworks benchmarking. SAPIEnS allows for the combined preprocessing and imputation steps assessment and optimal hyperparameters estimation that enhance different types of scATAC-seq data analysis, i.e. clustering, visualization and TF footprinting analysis.

SAPIEnS demonstrates that the optimal choice of imputation framework heavily depends on the task to solve and features (peaks) number. SCALE imputation with Boruta preprocessing improves clustering and visualization for small datasets. At the same time, large datasets clustering and visualization typically benefit more from Boruta preprocessing rather than imputation.

TFBS footprinting for HSC data is in concordance with clustering and may benefit from preprocessing with Boruta and imputation with either scOpen or SCALE. For this dataset, all imputation frameworks demonstrate simultaneous and proportionate



**Figure 4.** (A) A sample illustration for a strategy to estimate footprinting quality improvement for one of the transitions (HSC and MPP) for a specific imputation framework (SCALE + Boruta): for each of the cell populations, a difference of activity scores for each TF is calculated, ranked and inverted if needed, resulting in a TFRIS; for selected TFs, differences between ranks are summarized, resulting in a MRIS; (B) TFRIS for HSC to MPP branch for all imputation frameworks; (C) TFRIS for MPP and LMPP cell types for all imputation frameworks; (D) TFRIS for LMPP and CLP cell types for all imputation frameworks; (E) maximum TFRIS for all imputation frameworks; (F) Maximum MRIS for all imputation frameworks.



**Figure 5.** An ablation study of the footprinting pipeline for the HSC dataset: (A) an MLSI for peaks selected by preprocessing method and peaks matched to selected TF motifs; (B) an MLSI and maximum MRIS for peaks selected by preprocessing method; (C) an MLSI and maximum MRIS for peaks matched with selected TF motifs.

improvement in clustering metric values and digital footprinting results. This convergence and correspondence of analysis results derived from distant data modalities (peaks signal for clustering and open regions sequence structure for footprinting), as well as the fact that this convergence raises for a set of peaks associated with biologically relevant TFs compared with all observed peaks endorses the validity of the benchmarking results.

SAPIEnS has several limitations. For each imputation method, we had to select one optimal hyperparameter to avoid a combinatorial explosion. SCALE as a method based on a neural

network benefits vastly from selecting an optimal number of iterations to avoid under-training or over-training, whereas the imputation quality could be directly controlled by selecting the imputation rate.

In addition, scOpen has been originally developed to work on the whole scATAC-seq dataset. However, to compare the results of scOpen with SCALE, we had to use preprocessing methods. We observe that threshold filtering decreased scOpen results dramatically. However, the change of the preprocessing method has improved imputation metrics. In this way, scOpen could be applied to very sparse count matrices.

Preprocessing with Boruta, a feature selection algorithm, demonstrated improvement in multiple tests. Boruta searches for all relevant features that may be efficiently used for prediction rather than concentrating on finding a restricted group of features with the lowest classification error. However, to be applied for scATAC-seq data, it requires cell annotation. In this benchmark, we employ cell types annotation data based on the FACS sorting, which is not available for all the datasets. In these cases, other types of may be used, such as scRNA-seq co-embedding for cell types. Since Boruta is a general machine-learning method, any other relevant metadata on cell classes (gender, ethnicity, disease stage or stage of differentiation, mutation status, etc.) can be used for Random Forrest classification and subsequent feature selection making this approach much more flexible than Cicero. However, if no group annotation information can be fetched, Cicero preprocessing could be the method of choice significantly improving the results compared with a simple threshold method.

Independently considered results of TF footprinting have also limited validation power. We could not show improved metrics of footprinting for all of the known key TFs for HSC to CLP lineage, and some TFs (e.g. LYL1, FLI1, ERG1) do not demonstrate an expected pattern. Partially, this could be a result of the presence of such TFs only at the transient stage that has not been captured in the dataset due to a limited number of cells. The incomplete set of TFs might lead to a bias in footprinting metrics and affect the imputation performance.

A recent RNA-seq benchmark [7] has covered the effects of the imputation strategy on pseudotime analysis among others. In this work, we chose not to address this issue, given the fact that currently scATAC-seq data pseudotime-analysis is insufficiently developed and has very limited means for the analysis' results verification as compared with scRNA-seq data. Summing up, although all recruited validation approaches have limitations, in combination, given favorable convergence of their results, they deliver adequate verification power, suggesting that the strategy used in SAPIEnS may be used for scATAC-seq data imputation problem.

#### Key Points

- SAPIEnS is a benchmark for scATAC-seq imputation frameworks, a combination of state-of-the-art imputation methods with commonly used preprocessing techniques.
- SAPIEnS shows that the choice of the imputation framework heavily depends on the task to solve and the number of features (peaks) in the dataset.
- SCALE (Single-Cell ATAC-seq analysis via Latent feature Extraction) imputation with Boruta preprocessing improves clustering and visualization for small datasets while for large datasets either clustering or visualization do not typically benefit from imputation but only from preprocessing preferably with Boruta.
- Transcription factor binding sites (TFBS) footprinting is in concordance with clustering and may benefit from both preprocessing with Boruta and imputation with either scOpen or SCALE.
- A SAPIEnS database contains footprints of TFBS, and their activity scores in a multiple cell types.

## SUPPLEMENTARY DATA

Supplementary data are available online at <https://academic.oup.com/bib>.

## DATA AVAILABILITY

Breast Tumor data can be obtained from GEO at GSE112091. Forebrain dataset can be found at GSE100033. HSC dataset can be accessed at GSE96769. FibroCard dataset can be found by link [http://ns104190.ip-147-135-44.us/CARE\\_portal/ATAC\\_data\\_and\\_download.html](http://ns104190.ip-147-135-44.us/CARE_portal/ATAC_data_and_download.html). MouseAtlas data can be downloaded by link <https://atlas.gs.washington.edu/mouse-atac/>. T Cells data can be obtained at GSE107816. CellLines dataset can be accessed at GSE65360. PBMC5K dataset can be downloaded from [https://ftp.ebi.ac.uk/pub/databases/mofa/10x\\_rna\\_atac\\_vignette/seurat.rds](https://ftp.ebi.ac.uk/pub/databases/mofa/10x_rna_atac_vignette/seurat.rds).

## AUTHOR CONTRIBUTIONS STATEMENT

M.R. provided an idea of using preprocessing methods before scATAC-seq imputation methods. P.A. designed the architecture of the pipeline. L.S. conducted the experiments. P.A., Y.M. and A.S. analyzed the results. P.A., Y.M. and A.S. wrote and reviewed the manuscript.

## FUNDING

Ministry of Science and Higher Education of the Russian Federation (Grant No. 075-15-2022-310).

## REFERENCES

1. Buenrostro JD, Beijing W, Chang HY, Greenleaf WJ. Atac-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* 2015;**109**(1):21–9.
2. Furey TS. Chip-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet* 2012;**13**(12):840–52.
3. Hesselberth JR, Chen X, Zhang Z, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* 2009;**6**(4):283–9.
4. Vierstra J, Stamatoyannopoulos JA. Genomic footprinting. *Nat Methods* 2016;**13**(3):213–21.
5. Moyano TC, Gutiérrez RA, Alvarez JM. Genomic footprinting analyses from DNase-seq data to construct gene regulatory networks. *Methods Mol Biol* 2021;**2328**:25–46.
6. Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* 2020;**21**(1):22.
7. Hou W, Ji Z, Ji H, Hicks SC. A systematic evaluation of single-cell rna-sequencing imputation methods. *Genome Biol* 2020;**21**:1–30.
8. Li Z, Kuppe C, Ziegler S, et al. Chromatin-accessibility estimation from single-cell atac-seq data with scopen. *Nat Commun* 2021;**12**(1):6386.
9. Raevskiy M, Yanvarev V, Jung S, et al. Epi-impute: single-cell rna-seq imputation via integration with single-cell atac-seq. *Int J Mol Sci* 2023;**24**(7):6229.
10. González-Blas CB, Minnoye L, Papisokrati D, et al. Cistopic: cis-regulatory topic modeling on single-cell atac-seq data. *Nat Methods* 2019;**16**(5):397–400.



11. Kursa MB, Rudnicki WR. Feature selection with the boruta package. *J Stat Softw* 2010;**36**:1–13.
12. Pliner HA, Packer JS, McFaline-Figueroa JL, et al. Cicero predicts cis-regulatory dna interactions from single-cell chromatin accessibility data. *Mol Cell* 2018;**71**(5):858–71.
13. Xiong L, Kui X, Tian K, et al. Scale method for single-cell atac-seq analysis via latent feature extraction. *Nat Commun* 2019;**10**(1):4576.
14. Chen H, Lareau C, Andreani T, et al. Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome Biol* 2019;**20**(1):1–25.
15. Wang X, Lian Q, Dong H, et al. Benchmarking algorithms for gene set scoring of single-cell atac-seq data. *bioRxiv* 2023(01):2023.
16. Luecken MD, Büttner M, Chaichoompu K, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* 2022;**19**(1):41–50.
17. Liu Y, Zhang J, Wang S, et al. Are dropout imputation methods for scRNA-seq effective for scATAC-seq data? *Brief Bioinform* 2021;**23**(1):bbab442.
18. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. *ACM Comput Surv* 1999;**31**(3):264–323.
19. Omran MGH, Engelbrecht AP, Salman A. An overview of clustering methods. *Intell Data Anal* 2007;**11**(6):583–605.
20. Lorena AC, Garcia LPF, Lehmann J, et al. How complex is your classification problem? A survey on measuring classification complexity. *ACM Comput Surv* 2019;**52**(5):1–34.
21. Stupnikov A, Sizykh A, Budkina A, et al. Hobotnica: exploring molecular signature quality. *F1000Research* 2021;**10**:1260.
22. Jones KS. A statistical interpretation of term specificity and its application in retrieval. *J Doc* 2004;**60**(5):493–502.
23. Buenrostro JD, Ryan Corces M, Lareau CA, et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* 2018;**173**(6): 1535–48.
24. Patterson-Cross RB, Levine AJ, Menon V. Selecting single cell clustering parameter values using subsampling-based robustness metrics. *BMC Bioinformatics* 2021;**22**(1):1–13.
25. Stupnikov A, McInerney CE, Savage KI, et al. Robustness of differential gene expression analysis of rna-seq. *Comput Struct Biotechnol J* 2021;**19**:3470–81.
26. Zhang M, Yao C, Guo Z, et al. Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics* 2008;**24**(18):2057–63.
27. Stupnikov A, Tripathi S, de Matos R, et al. Samexplorer: exploring reproducibility and robustness of rna-seq results based on sam files. *Bioinformatics* 2016;**32**(21):3345–7.
28. Li Z, Schulz MH, Look T, et al. Identification of transcription factor binding sites using atac-seq. *Genome Biol* 2019;**20**:1–21.
29. Kulakovskiy IV, Vorontsov IE, Yevshin IS, et al. Hocomoco: towards a complete collection of transcription factor binding models for human and mouse via large-scale chip-seq analysis. *Nucleic Acids Res* 2018;**46**(D1):D252–9.
30. Chen X, Litzenburger UM, Wei Y, et al. Joint single-cell dna accessibility and protein epitope profiling reveals environmental regulation of epigenomic heterogeneity. *Nat Commun* 2018;**9**(1):4590.
31. Preissl S, Fang R, Huang H, et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat Neurosci* 2018;**21**(3): 432–9.
32. Wang H, Yang Y, Qian Y, et al. Delineating chromatin accessibility re-patterning at single cell level during early stage of direct cardiac reprogramming. *J Mol Cell Cardiol* 2022;**162**: 62–71.
33. Cusanovich DA, Hill AJ, Aghamirzaie D, et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* 2018;**174**(5): 1309–24.
34. Satpathy AT, Saligrama N, Buenrostro JD, et al. Transcript-indexed atac-seq for precision immune profiling. *Nat Med* 2018;**24**(5):580–90.
35. Buenrostro JD, Beijing W, Litzenburger UM, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 2015;**523**(7561):486–90.
36. 10X Genomics. 10k human pbmcs, multiome v1.0, chromium x <https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets>. 2021 (9 August 2021, date last accessed).
37. McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software* 2018;**3**(29):861.
38. Wolf FA, Angerer P, Theis FJ. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:1–5.
39. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008;**2008**(10):P10008.
40. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *the Journal of machine Learning research* 2011;**12**:2825–30.
41. Homola D, et al. boruta\_py. [https://github.com/scikit-learn-contrib/boruta\\_py](https://github.com/scikit-learn-contrib/boruta_py).
42. Aggarwal R, Lu J, Pompili VJ, Das H. Hematopoietic stem cells: transcriptional regulation, ex vivo expansion and clinical application. *Curr Mol Med* 2012;**12**(1):34–49.
43. Yoshida T, Ng SY-M, Zuniga-Pflucker JC, Georgopoulos K. Early hematopoietic lineage restrictions directed by ikaros. *Nat Immunol* 2006;**7**(4):382–91.