Single-cell ChIP-seq imputation with SIMPA by leveraging bulk ENCODE data

Steffen Albrecht [1,2], Tommaso Andreani [1,2], Miguel A. Andrade-Navarro [1], Jean-Fred Fontaine [1,*]

1 Institute of Organismic and Molecular Evolution (iOME), Johannes Gutenberg University Mainz, Mainz, Germany

2 Institute of Molecular Biology, Mainz, Germany

* to whom correspondence should be addressed: fontaine@uni-mainz.de

Abstract

Single-cell ChIP-seq analysis is challenging due to data sparsity. We present SIMPA (https://github.com/salbrec/SIMPA), a single-cell ChIP-seq data imputation method leveraging predictive information within bulk ENCODE data to impute missing protein-DNA interacting regions of target histone marks or transcription factors. Machine learning models trained for each single cell, each target, and each genomic region enable drastic improvement in cell types clustering and genes identification.

Body

The discovery of protein-DNA interactions of histone marks and transcription factors is of high importance in biomedical studies because of their impact on the regulation of core cellular processes such as chromatin structure organization and gene expression. These interactions are measured by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). Public data from the ENCODE portal, which provides a large collection of experimental bulk ChIP-seq data, has been used for comprehensive investigations revealing insights into epigenomic processes impacting chromatin 3D-

structure, open chromatin state, and gene expression to name just a few (ENCODE project consortium, 2012). Recently developed protocols for single-cell ChIP-seq (scChIP-seq) are powerful techniques that will enable in-depth characterization of those processes on single-cell resolution. ChIP-seq was successfully performed within single cells at the expense of sequencing coverage that can be as low as 1,000 unique reads per cell, reflecting the low amount of cellular material obtained from only one single cell (Rotem, Assaf, *et al.* 2015). Even though this low coverage leads to sparse datasets, scChIP-seq data was used to investigate relationships between drug-sensitive and resistant breast cancer cells; this would not have been possible with bulk ChIP-seq on millions of cells (Grosselin, Kevin, *et al.* 2019). Nevertheless, the sparsity of data from single-cell assays is a strong limitation for further analysis. In the context of ChIP-seq, sparsity means no signal observed for numerous genomic regions without the possibility to explain whether this is due to real biosample specific processes or to low sequencing coverage. Notably, sparsity may disable the investigation of functional genomic elements that could be of crucial interest. Hence, an imputation method is needed that completes sparse datasets from single-cell ChIP-seq while preserving the identity of each individual cell.

The first published imputation method for NGS epigenomic signals was ChromImpute (Ernst, Jason and Kellis, Manolis 2015), later followed by PREDICTD (Durham, Timothy J., *et al.* 2018), an improved method for the imputation of signal tracks for several molecular assays in a biosample-specific manner. The challenge of transcription factor binding site prediction was approached using deep learning algorithms on sequence position weight matrices (Qin, Qian and Feng, Jianxing 2017), and more recently by the embedding of transcription factor labels and k-mers (Yuan, Han, *et al.* 2019). Such methods show the successful application of machine learning and mathematical approaches in predicting epigenomic signals, however, their scope, being limited to either imputation of missing bulk experiments or sequence-specific binding site prediction, hampers their application to single-cell data. Imputation methods specialized for single-cell data are well established for RNA-seq, but the difference between RNA-seq and ChIP-seq data makes their application to scChIP-seq difficult. Recently, a method called SCALE (Xiong, Lei, *et al.* 2019) was published to analyze scATAC-seq (single-cell Assay for Transposase-Accessible Chromatin using sequencing) data, which is more similar to scChIP-seq. Although SCALE includes an imputation strategy, it was not yet applied and tested on scChIP-seq data.

Here we present SIMPA, an algorithm for **S**ingle-cell Ch**I**P-seq i**MP**ut**A**tion, and its validation on a scChIP-seq dataset of the H3K4me3 and H3K27me3 histone marks in B-cells and T-cells (Grosselin, Kevin, *et al.* 2019), currently the only available scChIP-seq dataset

for human cells. Different from most single-cell imputation methods, SIMPA leverages predictive information within bulk ChIP-seq data by combining the sparse input of one single cell and a collection of 2,251 ChIP-seq experiments from ENCODE. In order to better compare bulk and single-cell data, ChIP-seq regions (or significant signal/noise ChIP-seq peaks) are mapped to genomic bins (**Fig. 1A** and **Methods**). SIMPA's results for one single cell are obtained by using machine learning models trained on a subset of the ENCODE data related to a selected target. Derived from this target-specific subset, the classification features are defined by genomic regions detected in the single cell, while the class to predict is defined by a region observed in at least one target-specific bulk ENCODE experiment, but not in the single cell (**Fig. 1B**). In other words, by using this particular data selection strategy, SIMPA searches relevant statistical patterns linking protein-DNA interacting regions across single-cell specific regions of the target-specific ENCODE data for different cell types and the presence or absence of a potential region for the given single cell. SIMPA's machine learning models are able to use those patterns to provide accurate predictions (**Fig. 1C and S1**). Moreover, on the high-resolution H3K4me3 dataset, SIMPA achieved high recall rates (**Fig. S2**).

Next, we validated SIMPA in the task of separating B-cells from T-cells in the scChIP-seq dataset. Because of the better resolution available for H3K4me3 (processed as genomic bins of size 5kb), we present below results on this histone mark and refer to supplementary material for H3K27me3 (processed at 50kb). For benchmarking, we used an imputation method solely based on the single-cell dataset itself in contrast to SIMPA that takes advantage of information from a reference dataset. This reference-free imputation method is implemented by SCALE, an analysis method for single-cell ATAC-seq data (Xiong, Lei, *et al.* 2019). Furthermore, we implemented an average imputation method as a baseline approach (Schreiber, Jacob, *et al.* 2019). After applying a two-dimensional projection from a principal component analysis (PCA) on the sparse and imputed datasets, we observed that the separation between the cell types was drastically improved by SIMPA and by the reference-free method, contrary to the average imputation strategy (**Fig. 2A**).

Then, in order to validate the algorithmic concept of SIMPA we implemented two randomization tests in which either the ENCODE reference information was shuffled (Shuffled Reference) or the sparse single-cell input was randomly sampled (Randomized Sparse Input). Additionally, we applied SIMPA on the same data but with different histone marks as target. The selected histone marks were H3K36me3, a repressive mark functionally different to H3K4me3, and H3K9ac and H3K27ac, a group of two histone marks functionally related to H3K4me3. These two marks were used together to increase the

training data size. From this comparison, we observed that (i) the separation on the PCA projection is lost after removing statistical patterns through shuffling or randomization, (ii) separation quality stays moderate with an input mark functionally different to the real mark, and (iii) separation quality stays high using SIMPA with target histone marks functionally similar to the real mark (**Fig 2A**). Thus, the most relevant statistical patterns from the reference dataset are identified by both the selection of single-cell-specific regions and the selection of target-specific experiments. Similar observations were made for H3K27me3 although the separation between cell types was smaller than with H3K4me3, probably due to the lower bin resolution mentioned above (**Fig. S3**). The analysis of subgroups of B-cells and T-cells in the H3K4me3 data shows furthermore that the data structure was better preserved by SIMPA (**Fig. S4**).

Finally, we were interested to know whether enrichment analyses of cell-type-specific pathways for individual single cells can be improved after applying imputation. As H3K4me3 is an activating mark known to interact with promoters, we analyzed genes related to those promoters with the KEGG pathway analysis function of the Cistrome-GO tool (Li, Shaojuan, *et al.* 2019). As reported in **Fig. 2B**, the data was not enough within the sparse sets to show a significant pathway enrichment for any of the two cell types. Results from the reference-free strategy showed an improvement but not significant. However, with regions imputed by SIMPA, it was indeed possible to achieve significant enrichment scores and recover the cell-type-specific pathways for most of the cells. These results suggest that, contrary to the reference-free method that is limited to regions observed in the single-cell dataset, SIMPA is able to integrate functionally relevant information from the reference data in order to impute biologically meaningful regions.

In conclusion, the imputation strategy of SIMPA, as a novel approach in single-cell sequencing data imputation, is able to complete sparse scChIP-seq data of individual single cells, enabling better cell-type clustering and the recovery of cell-type-specific pathways.

References

ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome." *Nature* 489.7414 (2012): 57.

Grosselin, Kevin, et al. "High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer." *Nature genetics* 51.6 (2019): 1060.

Rotem, Assaf, et al. "Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state." *Nature biotechnology* 33.11 (2015): 1165.

Ernst, Jason, and Manolis Kellis. "Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues." *Nature biotechnology* 33.4 (2015): 364.

Xiong, Lei, et al. "SCALE method for single-cell ATAC-seq analysis via latent feature extraction." *Nature communications* 10.1 (2019): 1-10.

Durham, Timothy J., et al. "PREDICTD parallel epigenomics data imputation with cloud-based tensor decomposition." *Nature communications* 9.1 (2018): 1402.

Qin, Qian, and Jianxing Feng. "Imputation for transcription factor binding predictions based on deep learning." *PLoS computational biology* 13.2 (2017): e1005403.

Yuan, Han, et al. "BindSpace decodes transcription factor binding signals by large-scale sequence embedding." *Nature methods* 16.9 (2019): 858-861.

Li, Shaojuan, et al. "Cistrome-GO: a web server for functional enrichment analysis of transcription factor ChIP-seq peaks." *Nucleic acids research* (2019).

Schreiber, Jacob, et al. "A pitfall for machine learning methods aiming to predict across cell types." *bioRxiv* (2019): 512434.

Ghandi, Mahmoud, et al. "Enhanced regulatory sequence prediction using gapped k-mer features." *PLoS computational biology* 10.7 (2014): e1003711.

Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *Journal of machine learning research* 12.Oct (2011): 2825-2830.

Methods

ENCODE dataset preparation

To create the reference set that is used by SIMPA we downloaded all ChIP-seq experiments from the ENCODE portal that comply with the following criteria: the status is released, the experiment is replicated (isogenic or anisogenic), no treatment to the biosample, without genetic modification, and the organism is *Homo Sapiens* (human). For all experiments, we downloaded fully preprocessed sets of protein-DNA interacting regions as peak files: the *replicated peaks* for histone mark ChIP and the *optimal IDR thresholded peaks* for

transcription factor ChIP. If possible the peak files were downloaded for both assemblies hg19 and hg38 if one of the two was missing, we used the UCSC LiftOver tool to convert. Finally, we used 2251 experiments from different protein targets (antibody targets within the ChIP) and biosamples (either tissue or an immortalized cell line).

Data preprocessing

All reference experiments were converted from ChIP-seq peak sets to genomic bin sets. A bin is here a non-overlapping genomic region of a predefined size that can also be described by a unique identifier (ID), for example at a size of 5kb, the first bin is located on chromosome 1 from base 1 to 5000. We provide the reference data in bin sizes of 5kb and 50kb for hg38. Given one reference experiment, a bin is said to be “present” if there is at least one ChIP-seq peak that overlaps this bin, “absent” otherwise. In order to limit computational complexity, a reference experiment is finally described by a set of bin IDs.

SIMPA algorithm

SIMPA is an algorithm implemented in Python 3.7.3 for “**S**ingle-cell Ch**I**P-seq i**MP**ut**A**tion” that is applied to one single cell represented by a sparse set of scChIP-seq genomic regions (or peaks) provided by the user in bed format. Within the algorithm, the given single-cell bed file is converted into a set of bins SC describing the single-cell input. The user also provides the $target$ that is the name of the histone mark or transcription factor targeted by the antibody within the single-cell immunoprecipitation. The $target$ is needed to specify the training set that is composed of experiments from the ENCODE reference set.

Different from other single-cell imputation methods, SIMPA does not use information from other single cells. The imputation strategy is to use the sparse input from one given single cell to impute missing bins based on predictive information within bulk data between genomic regions bound by the target. In order to make the bulk data informative, first SIMPA collects all the ENCODE reference experiments available for the given $target$ that define the rows of the reference set matrix (*RS*) where columns represent bins:

$$RS = (a_{i,j}), 1 \leq i \leq n, 1 \leq j \leq m$$

with

$a_{i,j} \in \{0,1\}$describing a cell of the matrix with value = 1 when bin j in reference experiment i is present, 0 otherwise,

and where n is the number of experiments available for the given target, and m is the number of bins that are present in at least one of the $target$ specific experiments. As the rows are defined by the given $target$, the target-specificity is induced within this step.

Second, a subset of RS is created by selecting only the columns for bins that are present in SC to create the training features TF:

$$TF \subset RS,$$

$$TF = (a_{i,k}), 1 \leq i \leq n, 1 \leq k \leq s,$$

where k indexes a selection of bins from RS that are present in SC and with s the number of bins in SC (see Fig. 1B). At the same time, bins present in RS but not present in SC are collected and named as candidate bins c that are potentially imputed bins.

Third, SIMPA takes each candidate bin in c separately to compute an individual imputed probability ρ_i for each c_i. Given c_i, SIMPA trains a classification model cm_i based on TF defining the features and c_i as the class vector. Because an individual model is trained for each individual genomic bin, bin-specificity is induced to the whole approach. The imputed probability ρ_i is finally computed by cm_i that takes as input an artificial instance vector $a = (a_k), a_k = 1, 1 \leq k \leq s$. Consequently, ρ_i is the probability of c_i to be predicted for the imputed single-cell result, given the fact that all bins in SC are observed. As we use a Random Forest implementation from the scikit-learn (version 0.21.3) Python's library (Pedregosa, Fabian, *et al.*, 2011) with default settings to build classification models, the imputed probability is then the mean predicted class probability of the trees in the forest while the class probability of a single tree is calculated by the fraction of samples of the same class in a leaf.

Finally, SIMPA creates two files: one file in bed format and the other in SIMPA format described as a table listing the single-cell bins first, followed by the imputed bins sorted by the imputed probability. A line represents a bin described by its ID, its genomic coordinates, its frequency according to the target-specific reference experiments, and the imputed probability. Note, the first bins have no imputed probabilities as they represent the original sparse single-cell input (default value of -1 is assigned). However, the second file created by SIMPA is the imputed bed file containing the single-cell bins and bins with the highest imputed probability. The number of bins within this bed file is defined by the average number of bins present in the target-specific bulk experiments, 32584 for H3K4me3 (5kb bin size) and 12598 for H3K27me3 (50kb bin size).

Reference frequency and average imputation method

The reference frequency $freq$ of a particular bin j describes its presence in the reference experiments in RS:

$$freq_j = \frac{\sum_{i=0}^{n} a_{i,j}}{n}$$

where $a_{i,j} \in RS$ describes the values of bin j in experiment i.

The reference frequency, described above is, in general, a good indicator for the presence of a bin. Intuitively, the higher the frequency, the more likely the presence of the bin. We use this frequency to implement the baseline average imputation strategy. Hence, the average imputation strategy outputs the sparse bins from the single cell, plus imputed bins ranked by the reference frequency. The number of bins for the imputed result is the average number of bins observed for the target-specific reference experiment.

Cross validations

In order to validate whether machine learning models can be trained to accurately predict the observation of a bin we applied the following approach: given the target, 10 single cells were randomly sampled for both cell types (B-cell and T-cell); for each single cell the training feature matrix TF was created as explained for SIMPA while collecting also the candidate bins c. Then, for each candidate bin that defines the class vector, a Random Forest classification model was trained and evaluated by the area under ROC-curve within a ten-fold cross-validation. In addition, we used the area under precision-recall curve to better study the class vector imbalance.

Reference-free imputation

For reference-free imputation we used SCALE (a method developed for scATAC-Seq that uses the whole single-cell dataset for imputation; (Xiong, Lei, et al. 2019)) with default settings for the count matrices of H3K4me3 and H3K37me3 excluding gender-specific chromosomes but keeping all single cells. We used the flag *--binary* to receive a binary description of the presence of single bins within the imputed results for each single cell.

Randomization tests

We randomized the reference set obtained from the mark of interest from the bulk ENCODE data maintaining the same frequency observed for each bin. For this operation, we used the numpy shuffle function in Python (numpy 1.17.2). The second randomization was performed by simulating the single-cell bins as a set of random sequences with a similar amount, length and nucleotide distribution as observed for each single cell. For this operation we used the function “getNullseqs” from the package gkmSVM (Ghandi, Mahmoud, *et al.,* 2011) in R.

Similar and different mark imputation

To analyze the impact of the selected target that defines the training set, we analyzed the imputed results using histone marks with either similar or different functionality than the actual histone mark. For H3K4me3, an activating mark, the selected similar marks were H3K9ac and H3K27ac with 49 and 98 available experiments, respectively, and the selected different mark was H3K36me3 with 106 available experiments. For H3K27me3, we used H3K36me3 as similar, and H3K9ac and H3K27ac together as different marks. Using a collection of two similar marks (H3K9ac and H3K27ac) allowed us to increase the number of reference experiments used for training. The size of the reference set is 178 or 107 respectively when the real mark H3K4me3 or H3K27me3 is used as the target.

Preprocessing of scChIP-seq data (Grosselin et al.)

We downloaded the count matrices for H3K4me3 and H3K27me3 available in GEO under accession GSE117309 in 5kb and 50kb binning resolution, respectively. From the matrices, we derived bed files for every single cell excluding gender-specific chromosomes. SIMPA and other imputation methods were then applied on 25% of the single cells randomly sampled, 1520 bed files for H3K4me3 and 1128 bed files for H3K27me3.

Statistical analysis

In order to apply the Principal Component Analysis (PCA) implemented in scikit-learn, the bin sets of single cells were described by a matrix that has one single cell in a row and the bins described by the columns. The PCA was then applied to reduce this matrix to two dimensions using the default parameters.

Pathway enrichment analysis on sparse and imputed bin sets was performed using the *cistromego* tool downloaded from https://github.com/changxinw/Cistrome-GO and applied with default settings.

Implementation details and high-performance computing (HPC)

As described, SIMPA trains a Random Forest classification model specifically for each bin of each single cell. However, we observed that bins exist that have equal class vectors across the target-specific reference experiment set. Given a single cell and two or more bins with equal class vectors, SIMPA trains only one model for all of them resulting in equal imputed probabilities for these bins. Using this approach, the number of classification models was reduced on average from approximately 314,000 to 138,000 models trained for one single cell of the H3K4me3 data (5kb bin size).

In order to reduce run time spent on one single cell, SIMPA was implemented using an Open MPI interface for Python (mpi4py 3.0.2). The computationally heavy part of training classification models can be distributed to many CPU cores. We observed high CPU efficiency (> 97%) and 11 GB memory usage when using one full compute node (40 cores, 128GB RAM, Intel® Xeon® Processor E5-2630 v4) with a runtime of approximately 15 minutes for one single cell. We recommend using SIMPA within a cloud or high-performance computing system, if available. Considering the different validations in this paper based on the data for H3K4me3 with 5kb bins, we applied SIMPA on ~7,500 single cells for which 1.2 billion classification models were trained on the HPC system Mogon II (JGU, Mainz) within 24 hours (may be faster depending on the general workload of the system). Nevertheless, SIMPA can be applied on a standard computer. For one single cell, it took approximately 120 or 70 minutes using 2 or 4 cores, respectively, of an Intel® Core™ i5-4590 CPU @ 3.30GHz with 8GB RAM.

Data and Software Availability

The ENCODE data used by SIMPA is available on the GitHub page of the software in preprocessed format. Single-cell ChIP-seq data used for the validations was taken from GSE117309.

The SIMPA software is available at this link: https://github.com/salbrec/SIMPA

Acknowledgements

We would like to thank Pablo Mier for proofreading the manuscript. We would also like to thank the colleagues that tested SIMPA with their local machines: Kristina Kastano, Sweta Talyan, Jonas-Ibn Salem, and Gregorio Alanis-Lobato. Parts of this research were conducted using the supercomputer Mogon and advisory services offered by Johannes Gutenberg University Mainz (hpc.uni-mainz.de), which is a member of the AHRP (Alliance for High-Performance Computing in Rhineland Palatinate, www.ahrp.info) and the Gauss Alliance e.V. The authors gratefully acknowledge the Mogon supercomputer team. SA and TA would like to thank the International PhD Programme (IPP) of the Institute of Molecular Biology, Mainz, for financial support. We thank Susanne Gerber and Leszek Wojnowski for meaningful discussions.

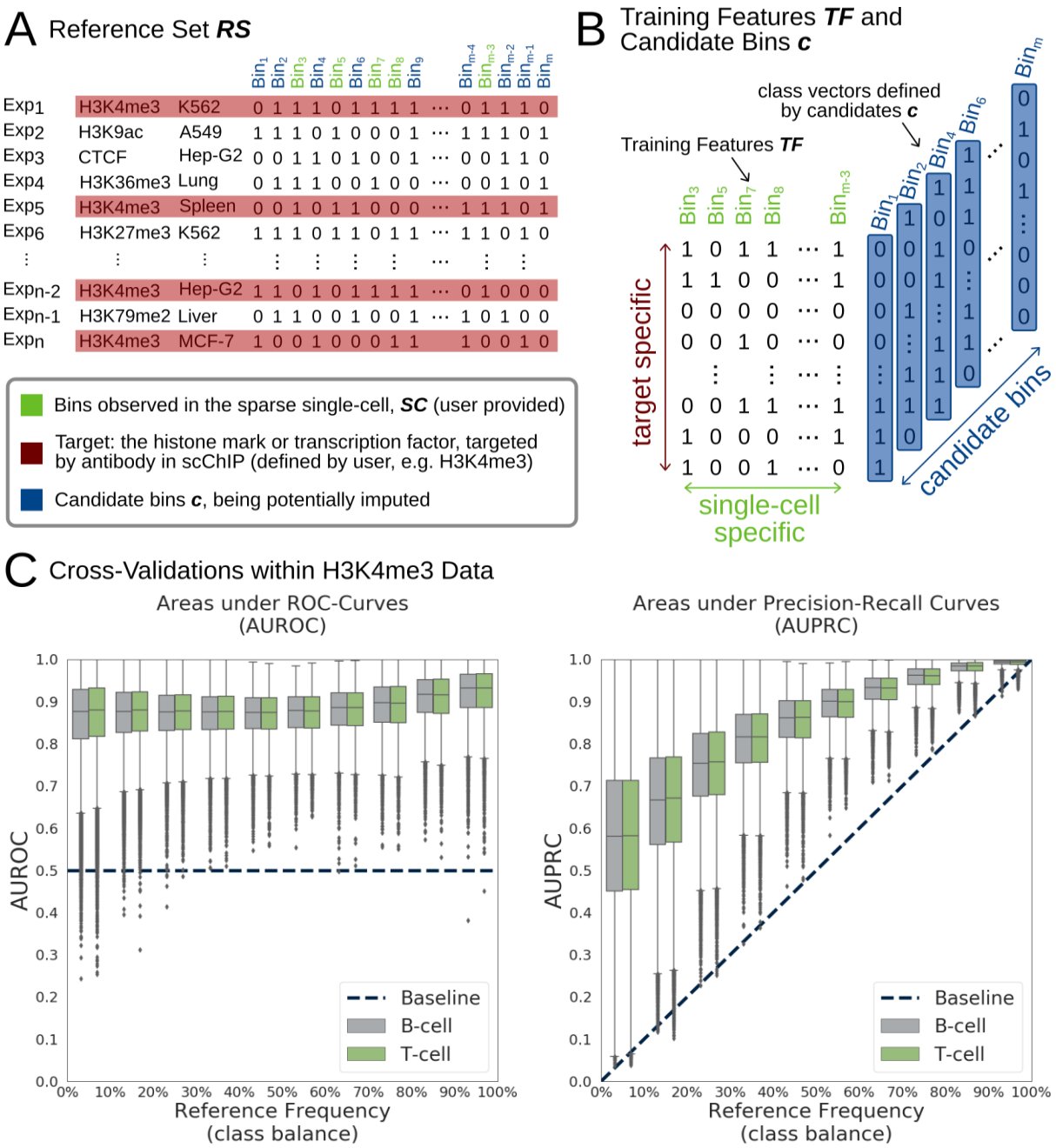
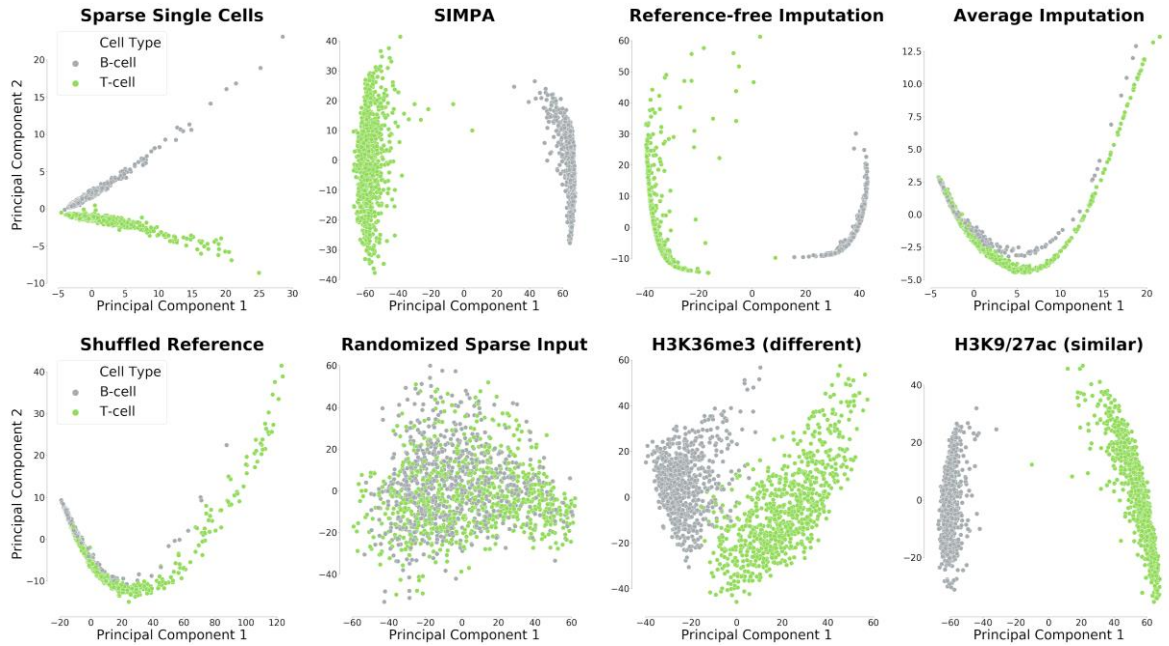


Fig 1. SIMPA's Algorithm and Cross Validations

A. Identified ChIP-seq regions from bulk experiments were downloaded from ENCODE and mapped to bins defined as non-overlapping and contiguous genomic regions of a defined length (5kb for H3K4me3 and 50 kb for H3K27me3) and covering the whole genome (the table). A bin is given a value of 1 for a particular experiment if there is at least one ChIP-seq region in this experiment that overlaps the bin, 0 otherwise. In total 2251 ChIP-seq experiments for several targets (histone marks or transcription factors) performed in several biosamples (tissues and cell-lines) were downloaded and preprocessed. Depending on the target specified by the user, the target-specific reference set ***RS*** is then created and contains

all experiments related to this target (red lines) and all bins observed for at least one of those experiments. **B.** The single-cell specific training feature matrix ***TF*** is created as a subset of ***RS*** by selecting only bins observed within the given single cell (green columns). All other bins from ***RS*** are the candidate bins (***c***; blue columns) and define the class vectors consisting of the corresponding values in ***RS***. For each candidate bin, a classification model is trained based on the training features and the class vector identifying associated experiments. **C.** Cross-validated evaluations of SIMPA's Random Forest performances to predict values of candidate bins in single cells within the H3K4me3 data. For each bin, a ten-fold cross-validation was applied and summarized as Area under ROC-Curve (AUROC) or Area under Precision-Recall Curve (AUPRC) (y-axes). Results for all bins are presented by boxplots subdivided by class balance in the candidate bins (percentage of "1" values in the bin) (x-axis). The dashed lines describe the baseline performance expected from a random classification model: 0.5 for AUROC and equal to the class balance for AUPRC.

A



B

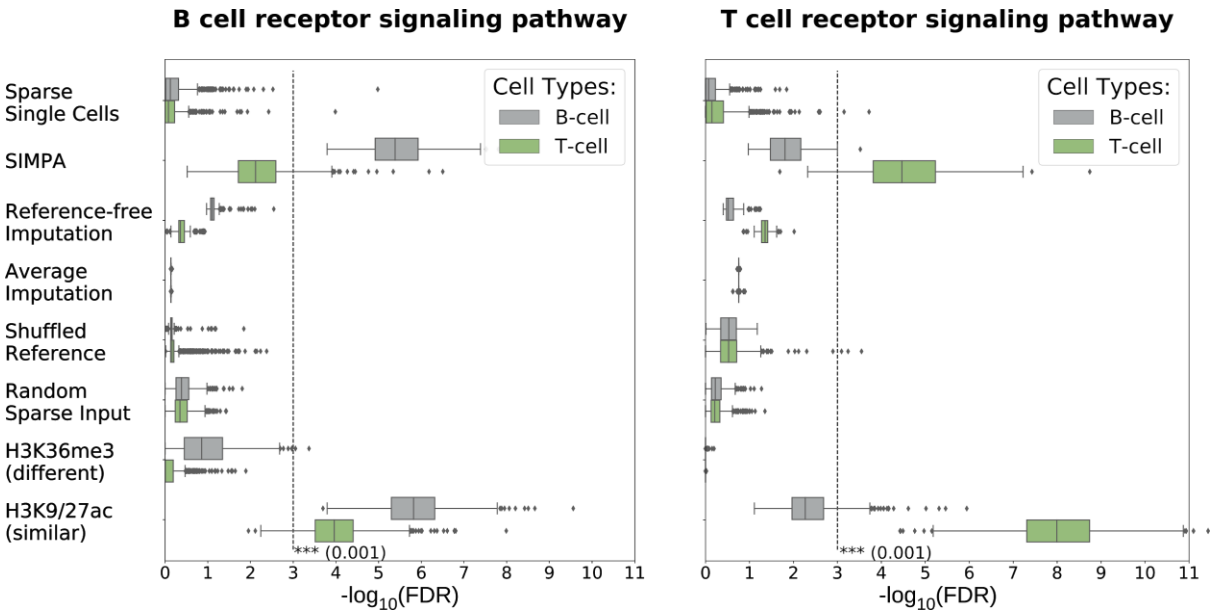


Fig 2. Cell-Type Specificity Validation

A. Separation of single cells according to cell type. Principal component analysis applied to the H3K4me3 data derived from the sparse single-cell data, SIMPA, reference-free imputation, average imputation based on expected frequencies in the reference set, shuffled reference set, randomized sparse input data, functionally different histone mark H3K36me3 as target instead of H3K4me3, and functionally similar histone marks H3K9ac and H3K27ac instead of H3K4me3. SIMPA achieves the best imputation by maximizing the separation of single cells (points) by cell types (colors) while showing no more artifacts due to data

sparsity (especially visible on the Sparse Single Cells plot). **B.** Pathway enrichment analysis. Boxplots show the significance of pathway enrichment analyses of genes annotated by single-cell regions as log-transformed false discovery rate (FDR; x-axis). Each dot represents the FDR of one single cell from the results of the different analysis experiments (y-axis). The dashed lines represent the log-transformed significance threshold of an FDR equal to 0.001. Only SIMPA achieves significant results by imputing preferably genomic regions associated with relevant pathway-related genes.