SPARTA imputes sparse single-cell ChIP-seq signals leveraging epigenomic bulk ENCODE data

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

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

2 Institute of Molecular Biology (IMB Mainz)

* fontaine@uni-mainz.de

ABSTRACT

Single-cell ChIP-seq analysis is challenging due to the sparsity of the data. We present SPARTA, a single-cell ChIP-seq data imputation method that leverages predictive information within bulk ENCODE data to impute missing protein-DNA interaction regions of target histone marks or transcription factors. By training hundreds of thousands of machine learning models specific to each target, each single cell and each region, SPARTA achieves high performance for clustering cell types and recovering regulatory elements specific for their cellular function.

The discovery of protein-DNA interactions from histone marks or 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 or gene expression. These interactions are measured by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). Public data as provided by the ENCODE portal that provides a large collection of experimental bulk ChIP-seq data, has been used for comprehensive investigations revealing insights to epigenomic processes with impact on chromatin 3D-structure, open chromatin state, or 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 could be used to investigate relationships between drug-sensitive and resistant breast cancer cells that wouldn't have been possible with bulk ChIP-seq on millions of cells (Grosselin 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 that there are numerous genomic loci without signal and for the vast majority of those it is not possible to explain whether these loci are not observed because of real biosample specific processes or because of the low sequencing coverage. Notably, sparsity may disable the investigation of certain functional genomic elements that could be of crucial interest. Hence, an imputation method is needed that completes sparse datasets from single-cell ChIP-seq to overcome the limitation while preserving the identity of each individual cell.

The first imputation method for epigenomic signals was ChromImpute (Ernst, Jason et al 2015) later followed by PREDICTD (Durham, Timothy J., et al 2018) that was also validated on more recent data with the goal to impute 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 Jianxing Feng, 2017), or more recently by the embedding of transcription factor labels and k-mers (Yuan, Han, et al. 2019). All these methods show the high potential of machine learning approaches and mathematical concepts for the prediction of 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 scRNA-seq, but due to the differences between RNA-seq and ChIP-seq data, it is difficult to apply these imputation strategies to scChIP-seq. Recently, another method named SCALE (Xiong, Lei, et al. 2019) was published to analyze scATAC-seq (single-cell Assay for Transposase-Accessible Chromatin using sequencing) data that is more similar to scChIP-seq data and includes also an imputation strategy. However, it was not applied and tested on scChIP-seq data.

Here we present SPARTA, an algorithm for **SPAR**se peaks impu**TA**tion for scChIP-seq, and its validation on a single-cell ChIP-seq dataset of the H3K4me3 and H3K27me3 histone marks in B-cells and T-cells. Different from most single-cell imputation methods, SPARTA 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 obtain bulk and single-cell data in the same format, ChIP-seq regions (or peaks) are mapped to genomic bins defined as small non-overlapping and contiguous genomic regions (**Fig. 1A and Methods**). SPARTA's results for one single cell are obtained by using machine learning models trained on a specified subset of the ENCODE data defined by target-related experiments, genomic regions detected in the single-cell, and ENCODE data of each region to calculate a probability for the imputation of regions that are not present in the single cell. (**Fig. 1B**). In other words, by using this specific data selection strategy, SPARTA searches relevant statistical patterns linking protein-DNA interacting regions across the target-specific ENCODE data for different cell types that explain the presence or absence of a potential region to be imputed for the given single-cell. SPARTA's machine learning models are able to use those patterns to provide accurate imputations (**Fig. 1C and S1**) while preserving the single-cell-specific information and data structure (**Fig. S2 and S3**).

We first validated SPARTA with a single-cell ChIP-seq dataset of B-cells and T-cells, consisting of two histone marks (H3K4me3 and H3K27me3) that were processed in bin format of size 5kb and 50kb, respectively (Grosselin et al. 2019). Because of the higher resolution available for H3K4me3, we present below results on this histone mark and refer to supplementary material for results on H3K27me3. For benchmarking, we used SCALE, an analysis method for single-cell ATAC-seq data that implements a different imputation strategy. Furthermore, as suggested by Schreiber et al. (Schreiber, Jacob, et al. 2019), we implemented an average imputation method as a baseline approach to be compared to more sophisticated concepts such as SPARTA or SCALE (**Methods**). After applying a two-dimensional PCA projection on the sparse data, we observed a good separation between the cell-types that was drastically improved by SPARTA and also SCALE, contrary to the average imputation strategy (**Fig. 2A**).

Next, in order to validate the algorithmic concept of SPARTA we implemented two randomization tests in which either the ENCODE reference information is shuffled (Shuffled Reference) or the sparse single cell input is randomly sampled (Randomized Sparse Input). Additionally, we applied SPARTA on the same data but with different histone marks as target input. The selected histone marks were H3K36me3, a repressive mark functionally different to H3K4me3, and H3K9ac and H3K27ac, a group of two histone marks more functionally related to H3K4me3. These two marks were used together to increase the feature space. From this comparison, we observed that (i) the separation on the PCA projection is lost after removing statistical patterns through randomization, (ii) separation quality stays moderate with an input mark different to the real mark, and (iii) separation quality stays high using SPARTA with an input mark that is more functionally similar to the real mark (**Fig 2B**). Thus, the most relevant statistical patterns from the reference dataset are identified by both the selection of single-cell-specific bins and the selection of target-specific experiments (see also **Fig. S4** for results with H3K27me3).

Finally, we were interested to know whether there were enough data available from single cells to find enriched cell-type specific pathways in annotations of bin-related genes. We applied the Cistrome-GO pathway analysis tool (Li et al. 2019) on the single-cell sparse bin sets and imputed bin sets from the different imputation methods and randomized tests. As reported in **Fig. 3** there was not enough data within the sparse bin sets to have a significant pathway enrichment for none of the two cell types. However, with bin sets imputed by SPARTA it was indeed possible to achieve significant enrichment scores and recover the cell-type-specific pathways. Results from SCALE showed as well an improvement but not as

significant as the results from SPARTA. These observations could be related to the potential of SPARTA's imputation strategy that allows to impute genomic regions from a larger scope defined by the bulk reference experiments compared to SCALE that is limited to the bins observed in the single-cell dataset.

In conclusion, the imputation strategy of SPARTA, as a novel approach in single-cell imputation, is able to complete sparse scChIP-seq data of an individual single cell, and enables 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 SPARTA we downloaded all ChIP-seq experiments from the ENCODE portal that comply with the following criteria:

- the status is released
- experiment is replicated (isogenic or anisogenic)
- no treatment to the biosample
- without genetic modification
- the organism is *Homo Sapiens* (human)

For all experiments, we downloaded fully preprocessed 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 USCS LiftOver tool to convert. In this way, we received 2251 experiments from different targets, the protein or protein variant targeted by the antibody within the ChIP, and biosamples that is either tissue or an immortalized cell line.

Data preprocessing

All reference experiments were converted into bin sets. A bin is a non-overlapping genomic region of a particular size that can also be described by a single ID, for an example at 5kb resolution, for the region on chromosome 1 from base 0 to 5000. We provide the reference data in the resolutions of 2kb, 5kb, and 50kb for hg38. Given one reference experiment, a bin is present if there is at least one peak that overlaps this bin. In order to decrease computational steps within SPARTA a reference experiment is finally described by a set of bin IDs.

SPARTA - overview - training - imputation

SPARTA is an algorithm for "**SPAR**se peaks impu**TA**tion" that is applied to one single cell represented by a sparse set of 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 (scChIP). 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, SPARTA is not using 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 that describe statistical relationships between genomic regions bound by the specific target. In order to make the bulk data informative, first SPARTA collects all the ENCODE reference experiments available for the given $target$ that define the rows of the reference set matrix (*RS*):

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

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 can be potentially imputed bins.

Third, SPARTA takes each candidate in c separately to compute an individual imputed probability ρ_i for each c_i. Given c_i, SPARTA trains a classification model cm_i based on TF defining the features and c_i as the class vector. In this way, bin-specificity is induced to the whole approach, because an individual model is trained for each individual bin that represents a genomic locus. 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. We use a RandomForest implementation from scikit-learn (Pedregosa, Fabian, et al., 2011) with default settings to build classification models.

Finally, SPARTA creates two files. One in bed format and the other in SPARTA format described as a table listing the single-cell bins first, followed by the imputed bins sorted by the imputed probability. One line represents a bin described by the bin-ID, its region, its frequency according to the target-specific reference experiments, and the imputed probability. Note, the first bins have no imputed probabilities as these are the bins from the sparse single-cell input and for those is assigned a value of -1. However, the second file created by SPARTA 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 (e.g. approximately 32000 for H3K4me3).

Reference frequency and impute probability

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 $a_{i,j} \in RS$ describing the the values ob bin j in experiment i.

The imputed probability is the probability of the bin calculated by the specific machine learning classification model that was trained within SPARTA for that particular bin. As

RandomForest is used, the imputed probability is 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.

Proof of the existence of predictive information

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 types (B-cell and T-cell); for each single cell the training feature matrix TF was created as explained for SPARTA 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 because the class vector can be imbalanced due to the reference frequency of a bin.

SCALE imputation

We used SCALE 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.

Average imputation

The reference frequency, described above is, in general, a good indicator for the presence of a bin. Intuitively, the higher this frequency, the more likely is its presence in a sparse bin set. 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.

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. The second randomization was performed by simulating the single cell bins as a set of random sequences with similar number, length and nucleotide

distribution as observed for each single cell. For this operation we used the function "getNullseqs" from the package gkmSVM (M. Ghandi 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. For H3K4me3, an activating mark, the similar selected marks were H3K9ac and H3K27ac with 49 and 98 available experiments resulting in a reference set composed out of 147. By using the collection of two similar marks, we were able to increase the number of reference experiments. Different to H3K4me3 is H3K36me3 with 106 available experiments. For H3K27me3, we used H3K4me3 as similar and H3K9ac and H3K27ac as different marks.

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. SPARTA 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 in two dimensions using the default parameters.

Pathway enrichment 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, SPARTA trains a Random Forest classification model specifically for each bin. However, we observed that bins exist that have equal class vectors across the target-specific reference experiment set. Given two or more bins with equal class vector SPARTA trains only one model for all of them resulting in equal imputed probabilities for

these bins as they are not different to each other. Using this approach, the number of classification models can be reduced from approximately 314.000 to 138.000 models trained for one single cell from H3K4me3 in 5kb resolution. SPARTA was implemented in Python 3.7.3.

In order to reduce run time spent on one single cell, SPARTA was implemented using an Open MPI interface. In this way the computationally heavy part of training the classification models can be distributed to many CPU cores. We achieved high CPU efficiency of > 97% when using one full compute node (40 cores, Intel® Xeon® Prozessor E5-2630 v4) that ran approximately 20 minutes for one single cell. We recommend to use SPARTA within a cloud computing or high-performance computing system, if available. Considering the core validations of this paper based on the data for H3K4me3 in 5kb resolution, we applied SPARTA on ~7,500 single cells for which 1,2 billion classification models were trained on the HPC system Mogon II that produced these results within 24 hours. Though, SPARTA can be applied on a local machine taking approximately 13 hours for one single cell using one core.

Code Availability

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

Data Availability

The ENCODE data used by SPARTA 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.

Acknowledgements

We would like to thank Pablo Mier for proofreading the manuscript. We would also like to thank the colleagues that tested SPARTA in their local machine: Jonas-Ibn Salem, Sweta Talyan 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 computing time granted on the supercomputer Mogon at Johannes Gutenberg University Mainz (hpc.uni-mainz.de). SA and TA would like to thank the International PhD Programme (IPP) of the Institute of Molecular Biology, Mainz, for

financial support. We thank Prof. Susanne Gerber and Prof. Leszek Wojnowski for meaningful discussions.

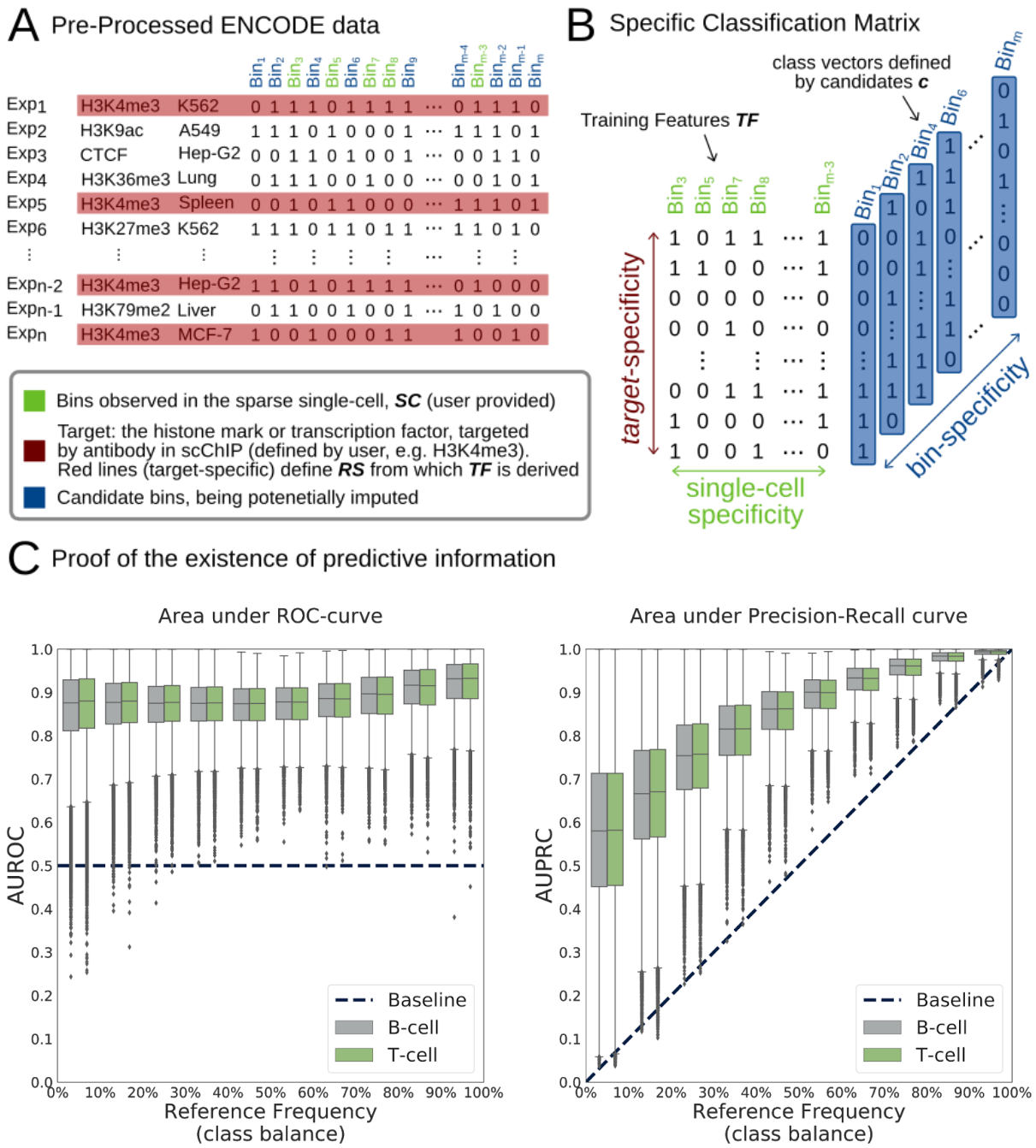

Fig 1 - Schematic overview of SPARTA

A ChIP-seq data 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). A bin is given a value of 1 for a certain experiment if there is at least one ChIP-seq peak of this experiment that overlaps the bin region, 0 otherwise. In total 2251 ChIP-seq experiments for several targets (histone marks or transcription factors) performed in several tissues and cell-lines were downloaded and preprocessed. According to the target, specified by the user, the reference set ***RS*** is created that contains only experiments for this target and bins observed for at least one of those

experiments. **B** In order to induce single-cell specificity, the training feature matrix ***TF*** is created as a subset of the ***RS*** by sub selecting only columns for bins observed within the given single cell (bin ID colored in green). The candidate bins (blue) are the bins that are potentially imputed (present in ***RS***, but not present in the single cell). For each candidate, a classification model is trained based on the training features and the class vector that is defined by the column containing the bin-specific values. **C** Proof for the existence of predictive information in target-specific (H3K4me3) reference set using single-cell specific bins for training and potentially imputed bins as class vectors. For a random selection of single cells and the target H3K4me3 the first steps of SPARTA were applied to receive the training features ***TF*** and candidate bins. For each bin, a ten-fold cross-validation was applied to evaluate the classification performance of Random Forest to predict its observation value 0 or 1. Bins, describing one sample for the box-plots were grouped by their frequencies in the reference set ***RS*** (x-axis) that describe the class balance of a classification task. The predictive performance of each classification task was evaluated by area under ROC-curve and area under precision-recall curve (y-axis). The dashed line describes the baseline performance expected from a random classification model. This baseline performance is 0.5 for the ROC-curve and dependent on the class balance in case of the precision-recall curve. (**Methods**).

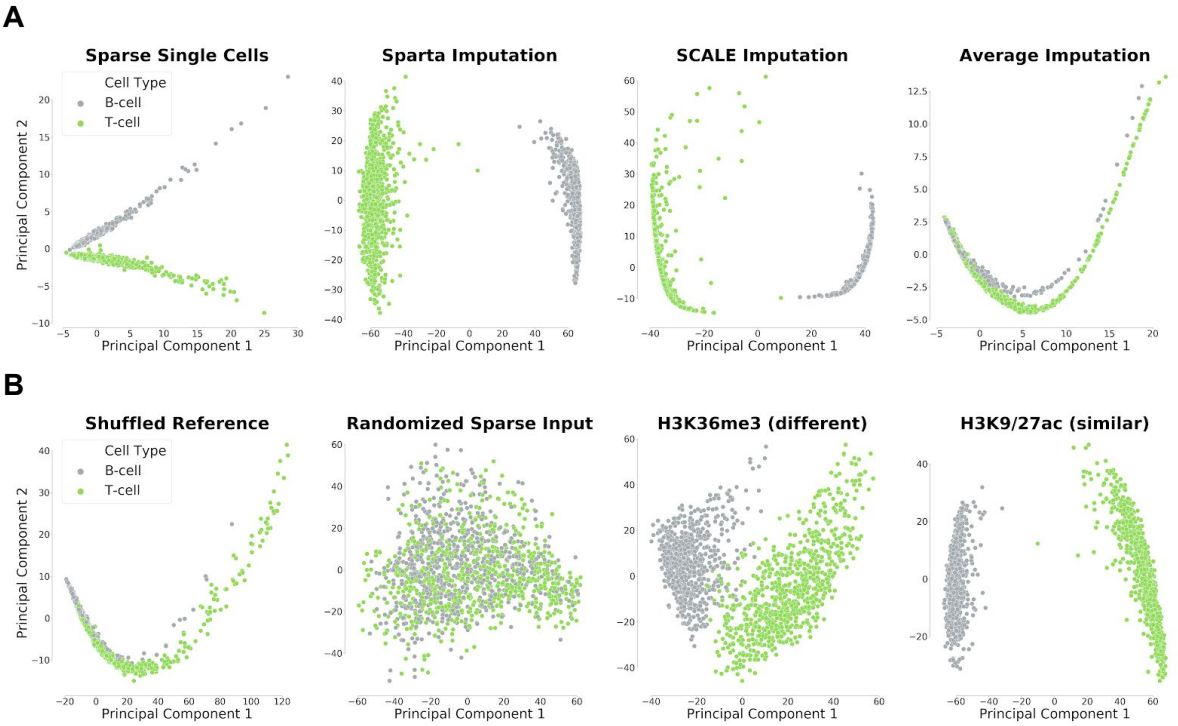

Fig 2 - Separation of single-cells according to cell-type

A+B Projection of the first two principal components of a PCA applied to the bin sets for single cells, derived from the sparse single-cell input data, different imputation strategies, and validation approaches. **A** Comparison between no imputation, Sparta, SCALE, and average imputation regarding the cell-type separation in the two dimensions described by the first two principal components. **B** PCA projections after applying randomization methods as a negative control to validate the algorithmic concept of Sparta followed by PCA projections from additional tests in which Sparta is used with histone marks that are different or similar to the actual histone mark used within the scChIP.

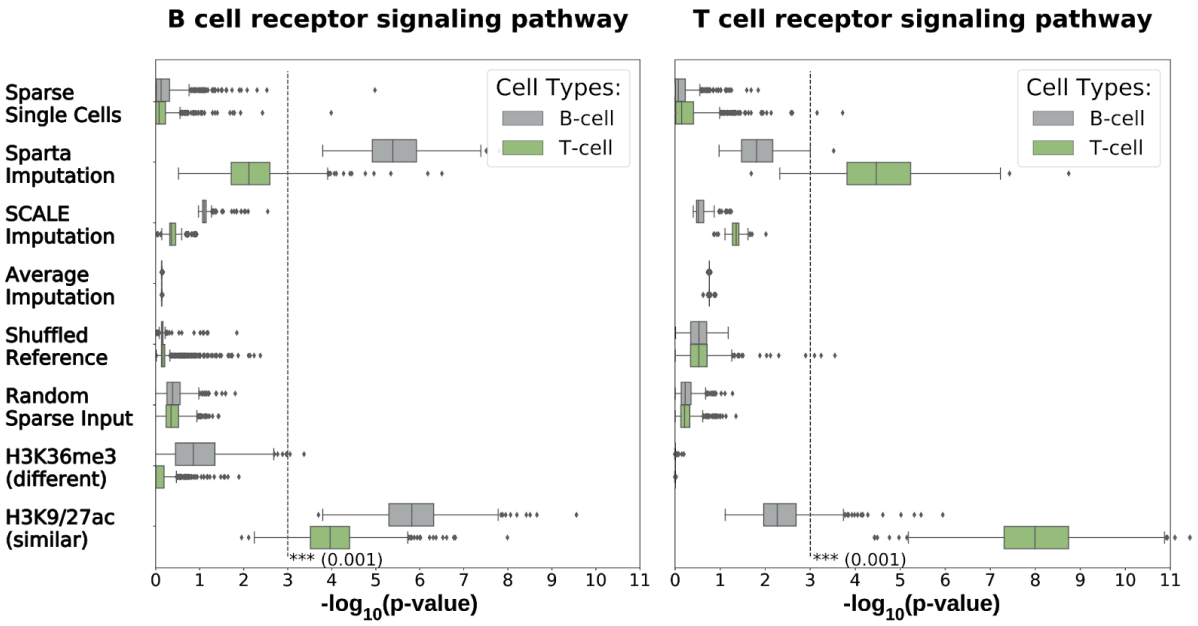

Fig 3 - Pathway and gene set analysis

Box-plots show the significance of a gene-set enrichment analysis for the specific pathways represented by log-transformed p-value (x-axis). Each dot represents the p-value of one single cell from the results of the different analysis experiments (y-axis). The dashed lines represent the log-transformed significance threshold of 0.001.