Random forest of Imputation Trees(RITS) for sparse single cell Genomics data

by

RACHESH SHARMA

Under the Supervision of

Dr Angshul Majumdar and **Dr Vibhor Kumar**

Indraprastha Institute of Information Technology Delhi

April, 2019

RITS for sparse single cell Genomics data

by

RACHESH SHARMA

Submitted
in partial fulfilment of the requirements for the degree of
Master of Technology

To

Indraprastha Institute of Information Technology Delhi

April, 2019

Certificate

This is to certify that the thesis titled '**RITS for sparse single cell Genomics data**' being submitted by RACHESH SHARMA to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

April, 2019

Dr. Angshul Majumdar
Department of Electronics and
Communication Engineering
Indraprastha Institute of Information
Technology Delhi
New Delhi 110 020

Dr. Vibhor Kumar
Department of Computational
Biology
Indraprastha Institute of Information
Technology Delhi
New Delhi 110 020

Acknowledgements

I take this opportunity to express my deepest gratitude towards Dr. Angshul Majumdar and Dr. Vibhor Kumar for helping me in bringing about this thesis. Their valuable instructions, inputs and brain storming sessions with them helped me in understanding the different facets of my thesis work. Their knowledge and expertise helped me in building the approach to carry forward my work in the right direction.

I would also like to thanks my parents and friends for encouraging me continuously through-out my research work.

RACHESH SHARMA
M.Tech.(CSE)
IIIT-Delhi

Abstract

A human body has billions of cells specialized with their own function and each cell carries genome in its nucleus. The activity of the genome is controlled by a multitude of molecular complexes called as epigenome. Previously scientists had a notion that human diseases are caused only due to changes in the DNA sequence or through the infectious agents present in the environment. However, recent studies have revealed that changes in the epigenome are also associated with disease.

Our aims is to create an imputation method for noisy, sparse and highly unbalanced single cell epigenome data. This problem is challenging as there is no imputation method for imputing huge and unbalanced dataset of single cell epigenome. Moreover, its analysis holds a significant amount of importance in the biological domain for preventing and curing many critical diseases.

Here we propose an imputation method called as RITs for imputing single cell epigenome profiles. We evaluated our proposed method through various possible techniques and compared its results with traditional imputation methods, although those imputation methods were made for imputing gene expression data. Our proposed method out-performs in every test and comes out as reliable imputation method even when we have huge unbalanced data. We tested our method on scATAC-seq dataset of cells from organs of the adult mouse to check the robustness and efficiency of this method. In all the conditions and tests, our imputation methods RITS remained at the top. The generality of RITs and it robustness for very noisy and sparse data-sets hints that it is the next generation imputation method for single cell profiles.

Contents

Certificate
Acknowledgement
Abstract
List of figures
List of tables

1. Introduction 10
   - 1.1. Types of Missing Data 10
   - 1.2. Techniques to handle Missing Data 10
   - 1.3. Need of next generation Imputation Method 12
   - 1.4. Current Method RITS 12

2. Problem Statement and Related Work 14
   - 2.1. MAGIC 15
   - 2.2. ScImpute 16
   - 2.3. DrImpute 16
   - 2.4. McImpute 17

3. Random forest of Imputation Trees (RITS) 18
   - 3.1. Methodology 18
   - 3.2. Imputation method used in RITS 23

4. Evaluation 24
   - 4.1. Recovery of missing data without over-prediction 24
   - 4.2. Detection of enhancers.....24
   - 4.3. Detection of Truly Differential Sites.....26
   - 4.4. Visualisation of data before and after Imputation 28
   - 4.5. Clustering Purity Test 34
   - 4.6. Cell Type Separability Test 35

5. Conclusion 37

Bibliography 38
Curriculum Vitae (CV) 41

List of Figures

Figure 1. RITS Complete Structure 19

Figure 2. RITS Phase 1 (Tree structure for hierarchal level 2) 20

Figure 3. RITS Phase 2 22

Figure 4. AUC Boxplot corresponding to cell type BJ and H1 25

Figure 5. AUC Boxplot for cell type BJ and H1 26

Figure 6. Willkoxon test between different cell types belongs to single cell epigenome dataset corresponding to five cell types.....27

Figure 7. Visualization single cell epigenome data for five cell types before and after imputation. 29

Figure 8. Visualization single cell epigenome data for 5 cell types before and after imputation using cosine similarity as distance function. 30

Figure 9. Visualization for ATAC-seq from BoneMarrow of adult mouse before and after imputation. 32

Figure 10. Visualization for ATAC-seq from liver of adult mouse before and after imputation. 33

Figure 11. Relative ARI Bar Plot over different scATAC-seq datasets. 35

List of Tables

Table 1. List of Imputation Methods 11

Table 2. Inter and Intra class correlation between GM128 and K562 cell type belong to single cell epigenome dataset. 36

Table 3. Inter and Intra class correlation between BJ and GM128 cell type belong to single cell epigenome dataset 36

1. Introduction

Missing data is a common problem in various research domains like medicines, recommendation systems, and climatic science [1-3]. There are multiple reasons for its emergence such as measurement error, mishandling of samples, deleting abnormal values or low signal to noise ratio. It makes a cause of concern for researchers because the analysis of missing data is arduous due to data ambiguity which affects the properties of statistical estimators such as mean, median, mode or variance and leads to a misleading conclusion. Therefore, the solution for it attracts a significant amount of attention in the research domain.

As according to "No Free Lunch Theorem" there is no algorithm which can solve every problem of a similar kind, therefore we tried to solve the missing data problem for single cell epigenome which belongs to biological research domain. We discussed more about single cell epigenome in chapter 2 and the proposed solution for its imputation in chapter 3.

1.1 Types of Missing Data

Missing data is mainly classified into three categories:

a. Missing Completely At Random (MCAR): When the missing value of a variable is entirely independent of known or unknown values of other observed variables. In other words, the missingness of data point is completely at random in data.

b. Missing At Random (MAR): When missing data of a variable can be inferred by known values of the same or other variable.

c. Missing Not At Random (MNAR): When missing data neither belongs to MCAR nor to MAR then that data counted in MNAR.

1.2 Techniques to handle Missing Data

The most natural solution to handle this problem is to remove observations having missing data, but it results in losing valuable information. A better solution is to predict missing values from an existing part of data, this

process of inferring missing values from data is called Imputation. There are various techniques proposed for the imputation. They are:

a. Mean Imputation: It is the most straightforward imputation method among all known methods. It consists of replacing unknown values for a given variable by the mean of known values of that variable. Its implementation is easy, and execution is fast but not good for big numerical datasets as it does not maintain any correlation with features.

b. K-nearest neighbors (KNN): This algorithm uses feature similarity to predict missing values. It finds the K-nearest neighbors for each sample and replaces the missing data of the given sample with averaging non-missing values of its neighbors. This algorithm is quite sensitive to outliers which is its main drawback.

c. Singular Value Decomposition (SVD): It is a straightforward matrix factorization method with mathematical proof. It is closely related to eigenvalues decomposition.

d. Multiple Imputation by Chained Equations (MICE): MICE is an iterative algorithm based on chain equation. It is far better than single imputation as it calculates the uncertainty of missing values in a better way. It uses a specified imputation model for each sample or variable and uses other samples or variables for its missing values prediction.

Imputation Method	**Category**
Expectation Maximization(EM)	EM
Iterative K-Nearest Neighbors	KNN
Least Squares	LS
Local-Least Squares	LS
Sequential Regression Trees	Tree
Bayesian Principal Component Analysis	SVD
Factor Analysis Model for Mixed Data	FA

Table 1: List of Imputation Methods

There are many other imputation techniques mentioned in Table 1, and most of them are characterized as one of the categories mentioned above.

1.3 Need of next generation Imputation Method

All the imputation methods have their own application and are highly effective over a particular kind of dataset. However, these imputation techniques cannot be applied directly over biological datasets because the downstream analysis of biological datasets generally need signal at cell type-specific sites and use of any imputation method will result in different biological meaning. Single cell epigenome profiles [5-9] which are a kind of biological data is a typical example of the very noisy, highly sparse and unbalanced dataset. The traditional methods proposed so far, have not performed imputation over such huge, noisy and sparse dataset. Although, there are few methods for imputing biological data-sets such as single cell gene expression profiles [13, 15], but their effectiveness for single cell epigenome data is questionable as it is more noisy and sparse than expression data. Therefore, there is a need for new generation imputation method which can identify similar cell types and cluster them correctly before performing any imputation. Hence we propose here an imputation method called RITS which stands for Random forest of Imputation Trees.

1.4 Current Method RITS

The study of epigenome gained importance since scientists discovered its role in regulation of expression of genes and its potential as target to prevent the development of dangerous disease like cancer. However, the analysis of single cell epigenome dataset is still challenging for biologists and scientists due to its sparse nature and absence of any imputation method for it. Hence, there is a need for an algorithm or method which can perform imputation over single cell epigenome and bring improvement in its downstream analysis.

While solving this problem, we found traditional imputation methods made for single cell expression data-set makes an error in classification and estimating drop-out rate. Since it is universally accepted that imputation depends on classification and wrong clustering can lead to non-optimal solutions, this motivated us to focus on enhancing the classification of cells

before performing an imputation. Therefore, we performed imputation over chains of clustering by ensembling them to avoid local minima for attaining a global solution. In our proposed approach, we generated ‘n’ numbers of trees and perform imputation operation individually in each tree. Finally, we use correlation based approach to combine the result of different imputation tree to get a single final imputed matrix. We named our approach Random forest of imputation trees (RITs), and benchmarked it’s performance for several scATAC-seq profiles.

2. Problem Statement and Related Work

Every living thing is composed of a wide variety of cell types as they are building blocks of organisms and every cell is unique in terms of structure with interdependent functions. It is essential to understand cells' behavior for the disease process and its proper treatment. It is crucial to learn genomics information for decoding characteristics, identity and functional status of each cell type for understanding development or disease process. Cells specialized for eyes are responsible for turning on genes that can detect light whereas red blood cell are specialised for oxygen transfer in the whole body. The activity of turning on or off of different sets of genes is controlled by epigenome. Study of epigenome became important when researchers found that human diseases are not only caused by the change in DNA sequence or infectious agent but also by change in the epigenome. The epigenome includes all the chemical compounds (Proteins, RNA) that surround DNA (genome) in a cell in such a way that they can regulate the activity of genes within the genome. The complexes attached to DNA which are responsible for modification of its function are often called as epigenomic markers. These marks are responsible for changing DNA instructions which are used by cells to express different phenotypes.

The study of the epigenome was mainly being done using bulk samples which typically have millions of cells. However, it does not help in finding the heterogeneity among cells in terms of their response to stimuli. Studying epigenome at single cell resolution becomes more crucial when single-cell RNA profiling is not able to differentiate two different cell with similar gene expression but different response. Moreover, single cell RNA-seq profiles cannot help in getting insight about non-coding genomic region activity. These issues, make researchers do profiling of single cell epigenome. Although histone modification and DNA methylation profiling for single cells have been performed but profiling of single cell, open chromatin is widely acceptable and used as it can reveal active and poised regulatory sites in the genome. In addition to this, it helps in determining the interaction pattern between chromatins.

Profiling of single cell open-chromatin needs peak calling as the first step after combining reads from multiple cells then estimation of the number

of reads lying on the peak for every cell needs to be done. Three protocol of profiling single cell open-chromatin has been developed, namely MNAse-seq, ATAC-seq and DNAse-seq whereas widely used protocol is scATAC-seq as its use is easy as compared to other protocols [6]. High drop-out rate and a large number of genuinely silent sites (true zeros) make scATAC-seq data sparse. Hence, imputation methods developed so far for single cell RNA sequence datasets [13,15,17] underperform over single cell open chromatin profile. There is a need for imputation method which can predict missing values for actual genomics sites which are not detectable easily (false zeros) without over-imputing true zeros. Such kind of imputation method would lead to the detection of enhancers specific to particular cell types as well as capture truly differentiable sites while comparing different cell types.

2.1 MAGIC

Profiling of scATAC profile is an emerging field of research, hence hardly any imputation method has been proposed till now for it, but there are imputation methods for single cell RNA-seq dataset which can be tried for predicting missing values in the scATAC profile. Although, for analyzing scRNA-seq data, methods are developed from a different perspective like identification of cell types, dimension reduction and clustering, these are also needed in scATAC signal in addition to some more peculiar analysis techniques. MAGIC is the first imputation method for predicting missing values in single-cell gene expression data by sharing information across similar cells using the concept of heat diffusion. It uses a graph structure to impute each cell in scRNA-seq data based on the weighted average of its neighbor while performing data smoothing and restore it to its underlying manifold. Moreover, it uses an adaptive Gaussian kernel which decreases similarity between two cells based on their distance exponentially. After imputation, each cell results in a unique expression vector because while averaging data across cells, each cell has a unique neighbourhood. MAGIC considers all zero counts as missing values which may introduce new biases and results in a blur out of genuine biological variation.

2.2 scImpute

Second imputation method which leverages the single cell expression data is scImpute. This method can identify dropouts and perform imputation only over those values to avoid entry of new biases to remaining data. Moreover, it uses clustering to enhance cell subpopulation which helps in detecting outlier cells and exclude them from imputation. This method aids the study of the dynamics of gene expression as well as improves the results for differential expression analysis.

It learns the probability of dropout in every cell based on a mixture model for each gene on the distribution of read-counts. It predicts missing values by using information of the same gene from similar cells. This method may not work successfully for the scATAC sequence in predicting missing values as the scATAC sequence is noisier as compared to expression data and leads to wrong clustering of sub-population of cells which is the prior task for imputation in scImpute method.

2.3 DrImpute

Both MAGIC and scImpute imputation methods have been developed for imputing single cell expression data. The bulk RNA-seq data does not allow profiling gene expressions at cell level variability. The bulk scRNA-seq have averaged gene expression of cells whereas scRNA-seq has gene expression at resolution of single cell. Moreover, dropouts in scRNA-seq have zero expression which is treated as missing values in other imputation methods. The dropout model of imputation is needed for scRNA-seq and DrImpute comes out as a solution to this problem.

As a first step, DrImpute identifies cluster based on similar cells. It performs imputation by averaging gene expression values of similar cells and this imputation is performed iteratively with different clustering results in achieving robustness before predicting final values. A notable point is that, zero expression values does not represents only missing value but also due to true biological reason. Even, it can be seen in imputation method published after this method that they have not considered all dropouts as missing value even while imputing bulk scRNA seq data.

2.4 McImpute

Scrutiny of scRNA-seq is advantageous for identifying cancer heterogeneity, new rare cell type and understanding of transcriptional changes that occur during development, but it becomes difficult due to insufficient input RNA and lowly expressed genes. Many imputation methods solved this problem by excluding lowly expressed genes. However such approach may not be the best solution for this problem as many transcription factors are sacrificed in this process.

Therefore, mcImpute comes out with a solution which does not drop lowly expressed gene. This algorithm models gene expression as a low-rank matrix and predicts missing values in the process of recovering the gene expression data by applying soft-thresholding iteratively on singular values of data. Moreover, mcImpute does not consider any assumptions of distribution for gene expression.

3. Random forest of Imputation Trees (RITS)

Imputation depends on the pattern that exists in missing data. The method which can correctly classify different cell types present in data, can predict missing values more accurately however correct classification of cells is a challenging problem. The reasons of wrong classification is highly unbalanced data and presence of noise. The imputation of single-cell RNA-seq is itself a difficult task when two different cell types have very few differences among them or when there is a minor cell type in the data. With single-cell epigenome data, the problems with imputation are more severe. For traditional imputation methods, the data is too challenging to handle as the number of genomic sites (features) in single-cell epigenome read-count matrix is too large and has high level of noise. Hence there is a need to make next-generation imputation technique for single cell epigenome datasets. We tried to resolve some of the challenging issues and came out with a more robust, memory and time efficient imputation method called RITS.

3.1 Methodology

It is universally accepted that accuracy of imputation methods is enhanced if correct grouping of similar cells is done. Hence, RITS uses a hierarchical approach of clustering for proper sub-cluster formation. Data pre-processing is an important step to start any computation over any dataset to achieve excellent results. Figure 1, shows the complete architecture of RITS. As shown in the figure, we first perform data pre-processing. In pre-processing step, we remove those peaks (sites or genes) which do not have a non-zero read count present in any cell. In addition to this, we did not consider those cells that are not active even for a single peak. We take a log transform of data after performing normalization over it. Let, the read count x_{ij} in a cell i on a site j having μ_i as a mean corresponding to cell i is

$$x'_{ij} = \log(x_{ij} / \mu_i + 1.01)$$

After data pre-processing, we have processed unimputed data which consist of neither cell nor gene having zero value throughout. Even then the degree of sparsity and noise in an epigenome data is high; hence

making a correct sub-clustering is not a trivial task. It made us opt semi-randomized approach of clustering in continuation with imputing. We create 'n' parallel trees to ensure our method does not trap in a local minimum. Computation corresponding to every tree is performed parallel in phase 1 without any interdependency on each other.

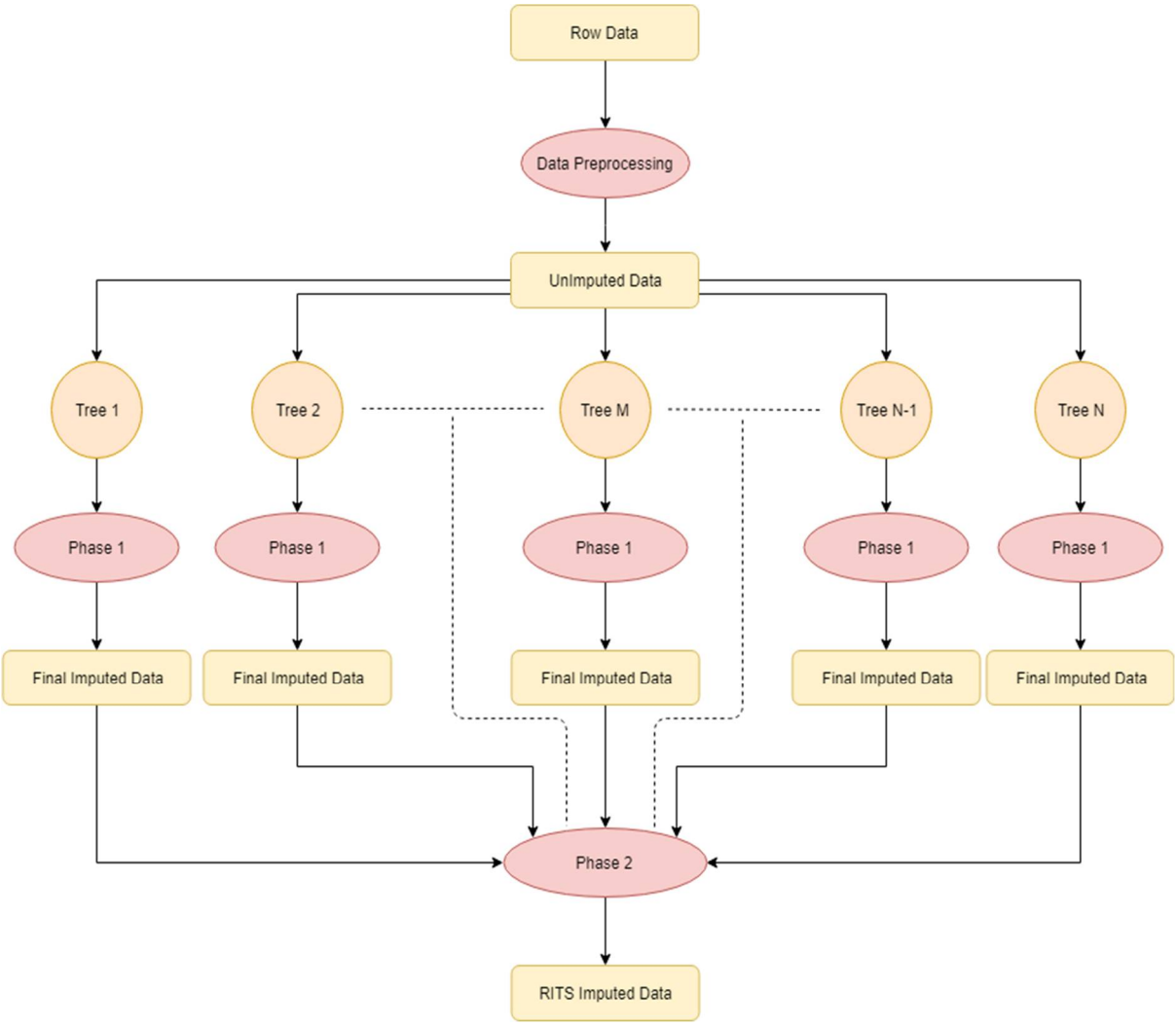


Figure 1: RITS Complete Structure

Steps corresponding to the computation of imputation for every tree in Phase 1 are the following:

Step 1: The depth of the tree is defined randomly within a range of 2 to max depth possible.

Step 2: Perform a preliminary imputation over complete unimputed data considering them in the same class.

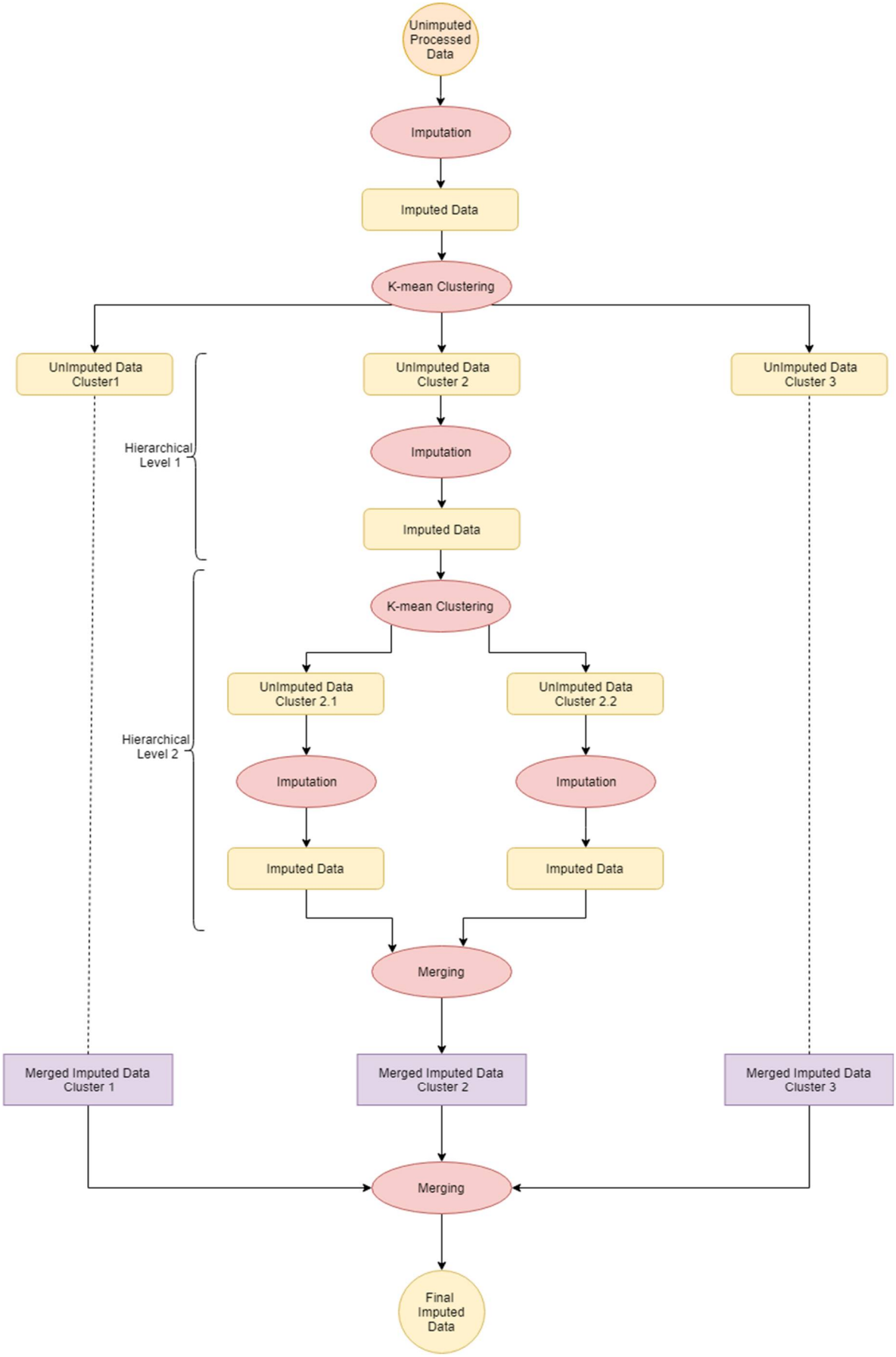


Figure 2: RITS Phase 1 (Tree structure for hierarchal level 2)

Step 3: Select m peaks randomly out of all peaks and perform SVD (Singular Value Decomposition) for dimensionality reduction over imputed data received as a result from Step 2 or 6.

Step 4: Perform k-means clustering after further reducing the dimension of dataset up to 3 to 5. The number of genomics sites and the value of k in k-mean is assigned randomly within a particular range based on condition.

Step 5: Divide samples from unimputed processed data according to clusters formed and remove those sites which have zero value for all cells or samples in that cluster.

Step 6: Perform imputation individually over clusters received from the previous step and repeat step 3 to 6 until the maximum depth of the tree is reached to find sub-classes correctly so that better imputation can be achieved.

Step 7: Assemble imputed matrices of all sub-classes received after step 6 by backtracking properly. Before assembling we add sites with zero values which are removed while imputing particular sub-class clusters in step 5.

The tree consists of multiple hierarchies and it is mandatory that the tree has at least two levels of a hierarchy. One level of an hierarchy completes on completion of steps 3 to 6 in phase 1 and it goes on increasing as the number of times iteration repeats from steps 3 to 6. Figure 2 represents tree having hierarchy level 2; it can be figured-out first imputation or imputation over complete unimputed data does not count in a tree hierarchy.

Once tree operation in phase 1 completes, we receive final imputed matrix corresponding to every tree running parallel. Now we need to compute the final imputed matrix and its computation is done in phase 2 as shown in Figure 3. The sequence of steps to follow in phase 2:

Step 1: Compute spearman correlation corresponding to every tree imputed matrix received after completion of phase 1 operations, with unimputed processed data.

Step 2: Extract top k most correlated either cells or peaks (it depends on the condition what user wants and how many i.e. value of k is defined by the user) from n final imputed matrices corresponding to every cell or peak present in the raw data.

Step 3: Assign an imputed value to the matrix cell belongs to the i^{th} gene and j^{th} cell type from the maximum imputed value belonging to tree extracted from top k corresponding to that cell or gene.

We get RITS imputed matrix as a result of step 3 of phase 2 and this complete process is RITS overview. We can compute imputation on large single cell data-sets (i.e. the data whose dimension is too high, the other imputation methods not only failed to impute that data correctly but also failed to even impute it.) making it strongest among all other imputation methods. RITS can be computed over highly bulk data by dividing them randomly into small pieces and then performing RITS method imputation separately before combining them again to achieve final impute RITS matrix.

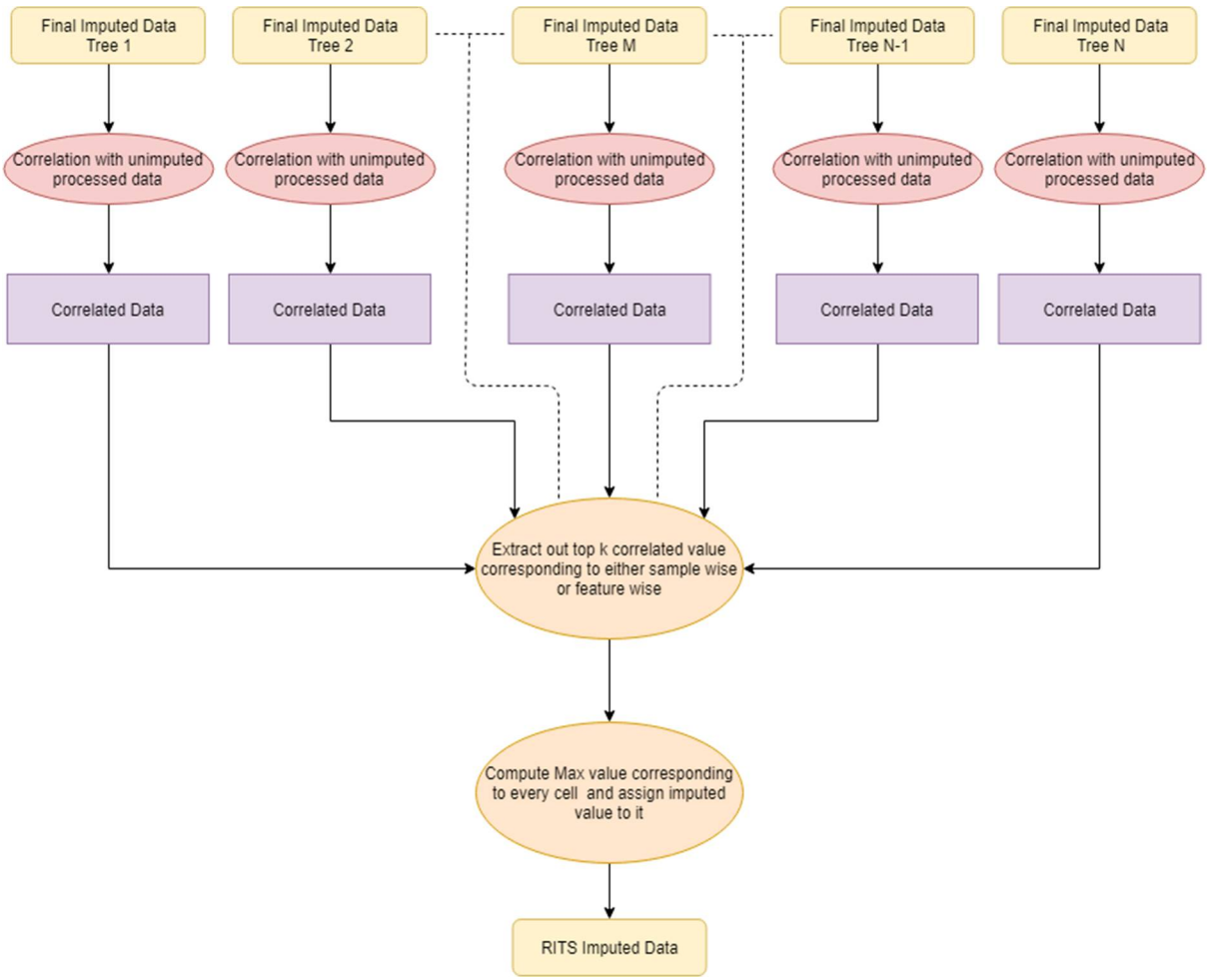


Figure 3: RITS Phase 2

3.2 Imputation method used in RITS

RITS is robust enough to handle error caused due to wrong classification and imputation but it is worth to discuss its baseline program which uses nuclear norm minimization approach. Given a read count matrix Y consisting of genomic-peaks as column and cells or sample as a row. The problem of imputation is to recover ideal matrix X through the observed matrix Y which is a mask version of the true matrix X such that

$$Y = A(X) \quad \ldots(1)$$

Here, A is a masking operation operator which causes sub-sampling and assigns zero where elements are not observed otherwise assigns one. In order to solve this problem with matrix factorization, one should know approximate rank r of matrix X, therefore researchers solve this problem directly with a constraint that solution has low-rank which is NP hard problem. Hence, nuclear norm minimization [10] is proposed for relaxing NP hard problem to its convex surrogate

$$\min_X ||X||_* \text{ such that } Y = A(X) \quad \ldots(2)$$

Here $||.||_*$ is nuclear norm calculated as the sum of singular values present in matrix X. This does not solve the complete problem, it is just a relaxed version of problem. Hence Lagrange multiplier (λ) is introduced as:

$$\min_X || Y - A(X) ||^2_F + \lambda||X||_* \quad \ldots(3)$$

This equation needs to be solved iteratively. Now, it uses Majorization-Minimization (MM) approach for solving the minimization problem and now equation (3) will be:

$$\min_X || B - X ||^2_F + \lambda||X||_* \quad \ldots(4)$$

Here, $B_{K+1} = X_K + (1/a)\, A^T (Y - A(X_K))$ and X_K is the matrix at Kth iteration and 'a' is a constant parameter. As inequality suggests $||P_1 - P_2||_F \geq ||s_1 - s_2||_2$, where s_1 and s_2 are singular value of matrix of P_1 and P_2 respectively. The problem is now solved using this rule instead of directly solving minimization problem (4).

$$s_X = \text{sign}(s_B) \max(0, |s_B| - \lambda/2) \quad \ldots(5)$$

4. Evaluation

Imputation decreases noise and sparsity of data; even then one has to evaluate and compare its result with others for finding out its efficiency in improving the profiling of any biological data. Therefore, we perform systematically different experiments for the evaluation of RITS performance.

4.1 Recovery of missing data without over-prediction

We evaluated RITS over dataset published by Buenrostro et al.(2015) of scATACseq profiles corresponding to 5 cell types namely GM12878, K562, HL60, BJ and HESC having 1622 cells and 92447 genes. We first evaluated whether RITS imputation enhances the quality of single cell epigenome data corresponding to scATAC-seq protocol. We tried to find in which of the five cell type for every genomic site present in imputed scATAC-sec dataset overlapped with sites of bulk ATAC-seq. We calculated ROC-AUC for each cell to estimate the coverage of true peaks present in bulk ATAC-seq in respective cell types. In addition to this, we measured the accuracy of detecting true zero as imputation may result in over-prediction which defeats the purpose of single cell open-chromatin profiling. For detection of true zeros and true peaks RITS outperformed other tools like MAGIC and scImpute. Figure 4 shows boxplot corresponding to different cell types present in a single cell open-chromatin profile, and it can be easily verified that RITS not only outperformed compared to others but also improved performance over base imputation program. Imputation efficiency depends on classification and randomization in selecting sites at each hierarchy level in RITS enhancing sub-clustering accuracy which leads it to achieve excellent results. Here, we have compared RITS with the ideal cases of scImpute, MAGIC and base imputation methods.

4.2 Detection of enhancers

Studying the activity of regulatory element like enhancers with cell-type specific activity, is one of the reasons for profiling of open chromatin.

Enhancers are a small region of DNA where transcription factor bond to activate and enhance the transcription of a particular gene. Using open-chromatin profiles, researchers often predict enhancers using the technique of highlighting for specific cell type activity. We adapted this technique to highlight cell type activity for the genomic site by dividing the read-count matrix of open chromatin data with its average across all the cells.

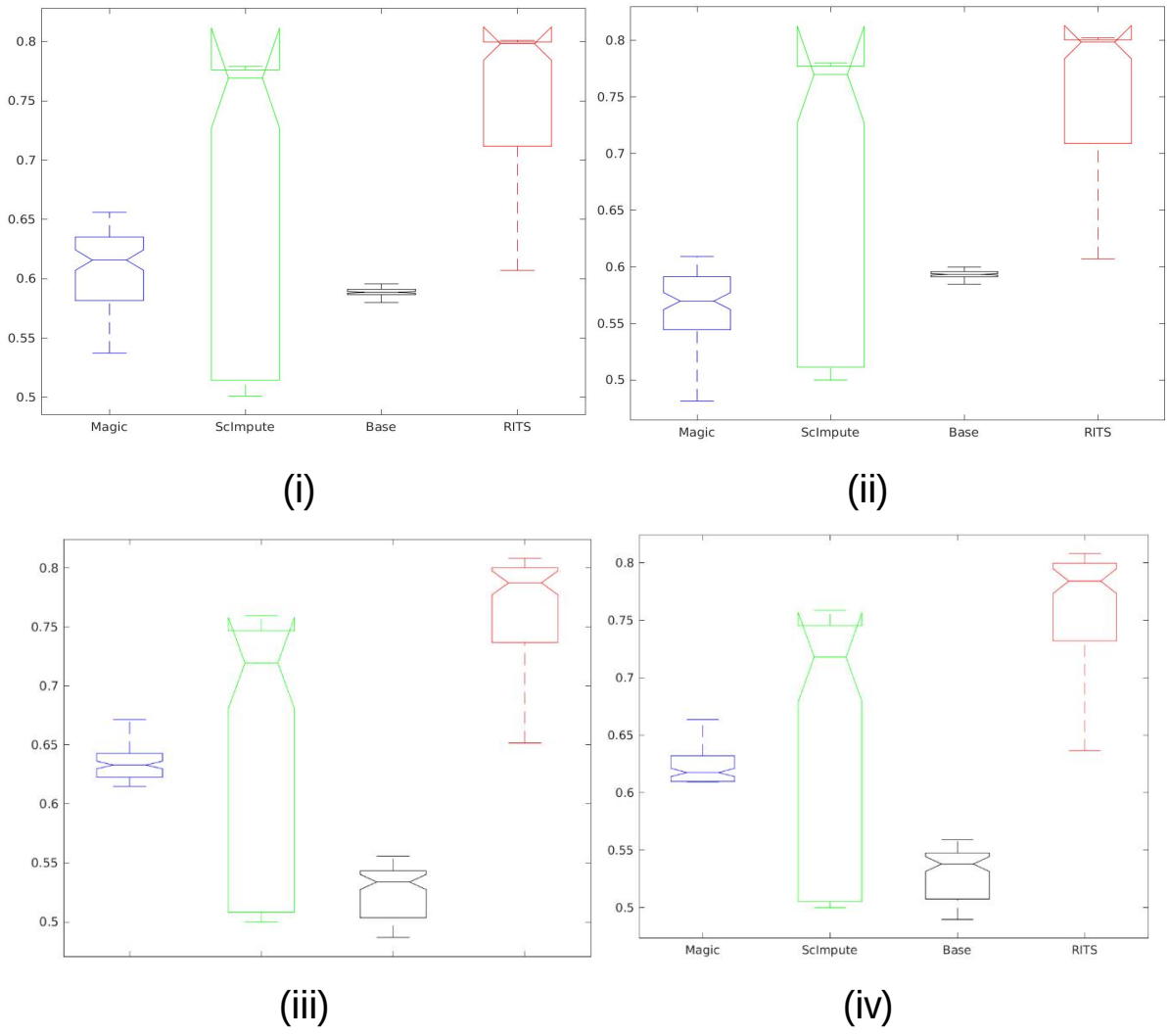
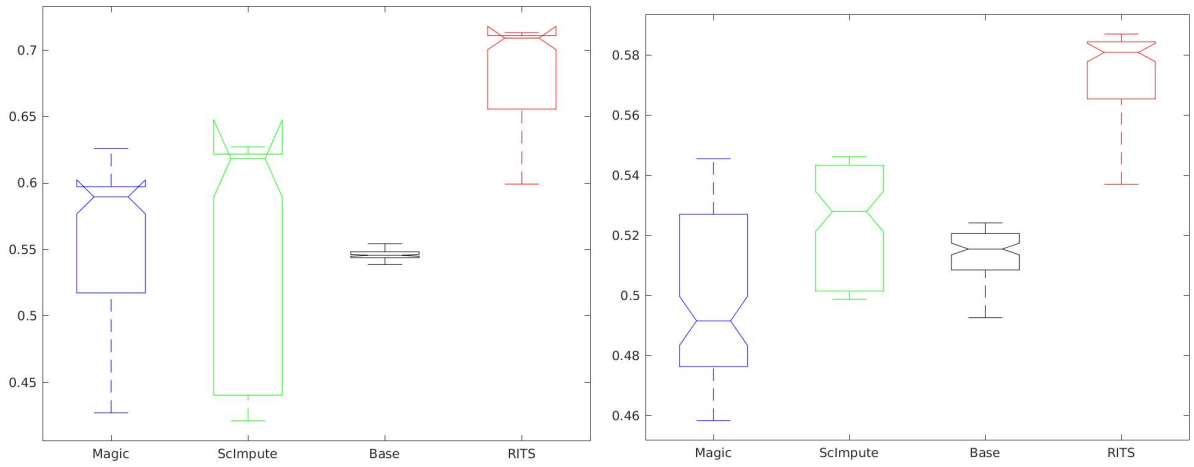


Figure 4: AUC Boxplot corresponding to cell type BJ and H1 where figure (i) represents true peaks AUC for BJ cell , figure (ii) represents true zeros AUC for BJ cell, figure (iii) represents true peaks AUC for H1 cell and figure (iv) represents true zeros AUC for H1 cell.

We evaluated the detection of enhancer in single cell epigenome imputed data and RITS provides higher coverage as compared to other methods for enhancers. Figure 5 represents boxplot obtained for enhancers in single cell epigenome data corresponding to different cell types. Here, we computed ROC-AUC for every cell type as enhancer, either active or inactive for particular sites corresponding to specific cell type. Results will depict the effectiveness of imputation that it can be easily categorized into on or off state in cells.



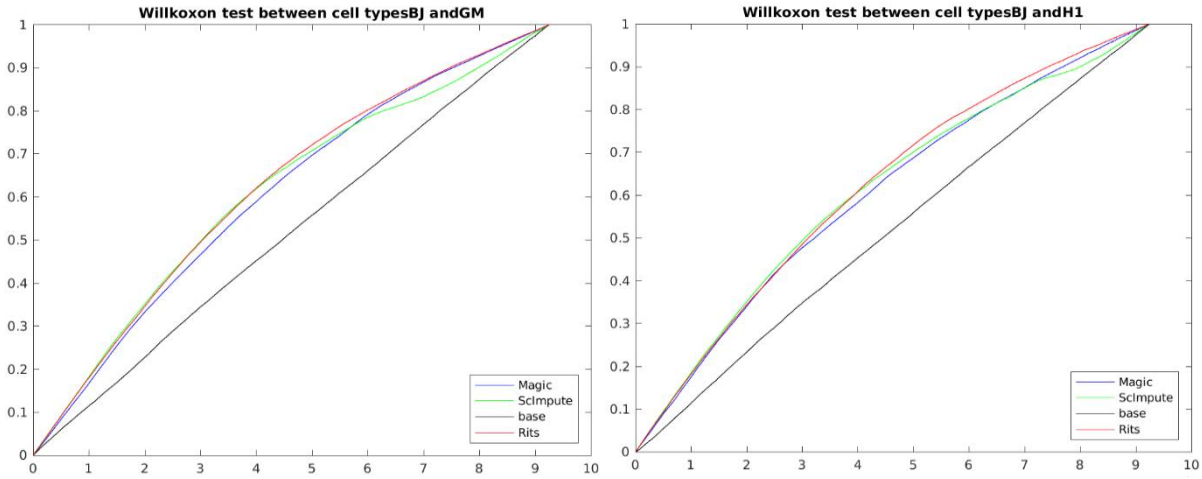
(i) (ii)

Figure 5: AUC Boxplot for cell type BJ and H1 when enhancer is active are shown in figure (i) and figure (ii) resp.

4.3 Detection of Truly Differential Sites

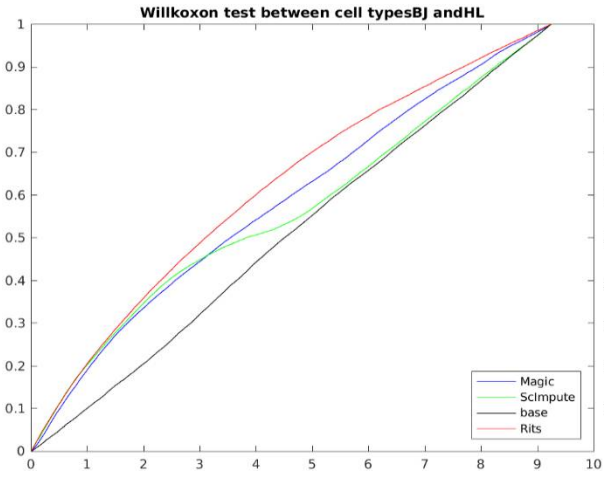
Cells may look similar in terms of gene-expression but in terms of epigenome profile they might be different. Detection of differentially active genomics sites helps in revealing differences among poising and activity of regulatory elements in ATAC-seq profile occurs due to environmental changes or disease(s). Moreover, if the cells can differentiate using genes, then it shows imputation performed over dataset is very effective and efficient.

In this test, we don't have a true difference ground truth available for different combinations of cell types. Hence, we give importance to those

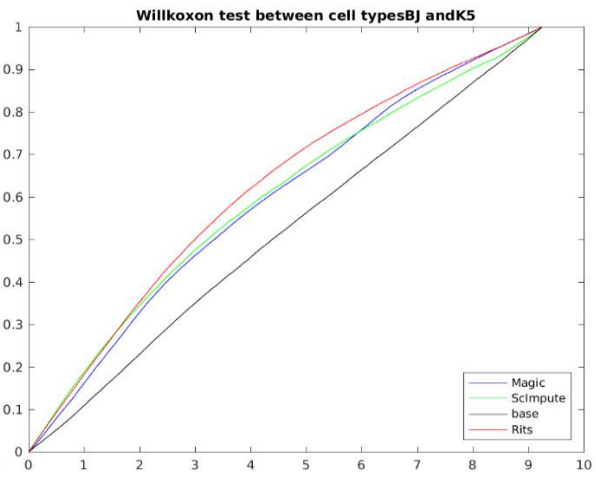


(i)

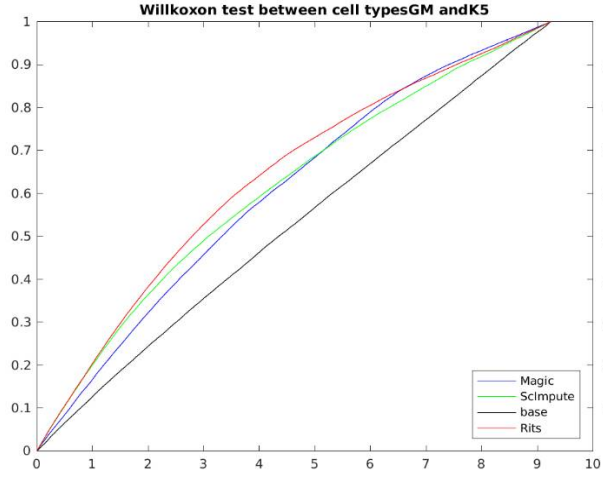
(ii)



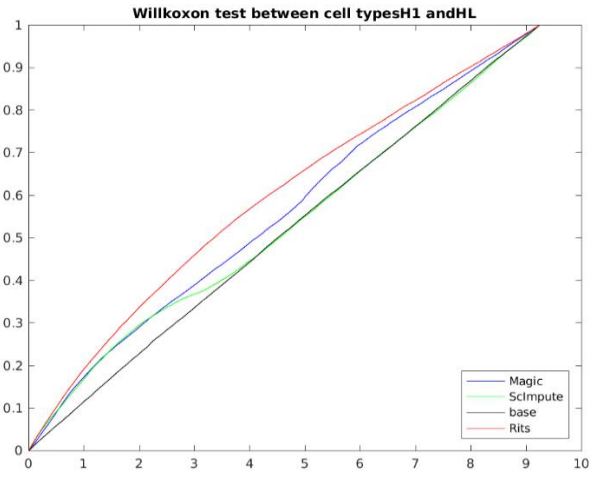
(iii)



(iv)



(v)



(vi)

Figure 6: Willkoxon test between different cell types belongs to single cell epigenome dataset corresponding to 5 cell types.

genes while performing test over the combination of two cell types which shows different behaviour for them. We outperform even in this evaluation figure 6 shows the Willkoxon test over six combinations of cells out of 10 possible combinations corresponding to single cell epigenome dataset having five cell types. We have shown results in figure6(v) between GM 128 (GM) and K562 (K5) cell types as their number of samples is large in dataset. We have also shown this test result in figure 6(ii) between BJ and H1 cell types as they have lowest number of cells present in complete dataset. For both whether number of cells corresponding to particular group is low or high, RITS always performed best. In addition to this, we even compute over unequal division samples like BJ and GM, BJ and K5 etc.

4.4 Visualisation of data before and after Imputation

Visualization is always one of the important ways for evaluation and analysis of any data and result of the operation performed over it. It plays a crucial role in analyzing single-cell open chromatin profile to understand separability obtained among different cell types present, after the imputation as chromatin profile consists of sparse and noisy data in a high percentage. Reduction of dimension is one of the primary tasks to visualize and analyse single cell open chromatin profile. The dimensionality reduction can never be an easy task especially when there is a high degree of noise and sparsity which always becomes an obstacle while clustering and classification.

For visualisation, we performed tSNE over imputed as well as non-imputed data with distance function 'Euclidean'. Figure 7 represents the tSNE plot for scATAC-seq corresponding to five cell types. The tSNE plot over un-imputed data shown in figure 7(i) and one can easily interpret that performing classification and then imputation is not a cup of a tea for any imputation method. MAGIC and scImpute are far-away in comparison with RITS imputed results as figure 7(ii) and figure 7(iii) shows, these imputation methods failed to classify correctly. Although base imputation method shown in figure 7(iv) has quite good results but it separated out K562 cell types into two part with huge distance which is not a case in RITS i.e. figure 7(v). All these imputation performed in their ideal condition except the imputation through RITS. In the real world scenario, biologists don't know

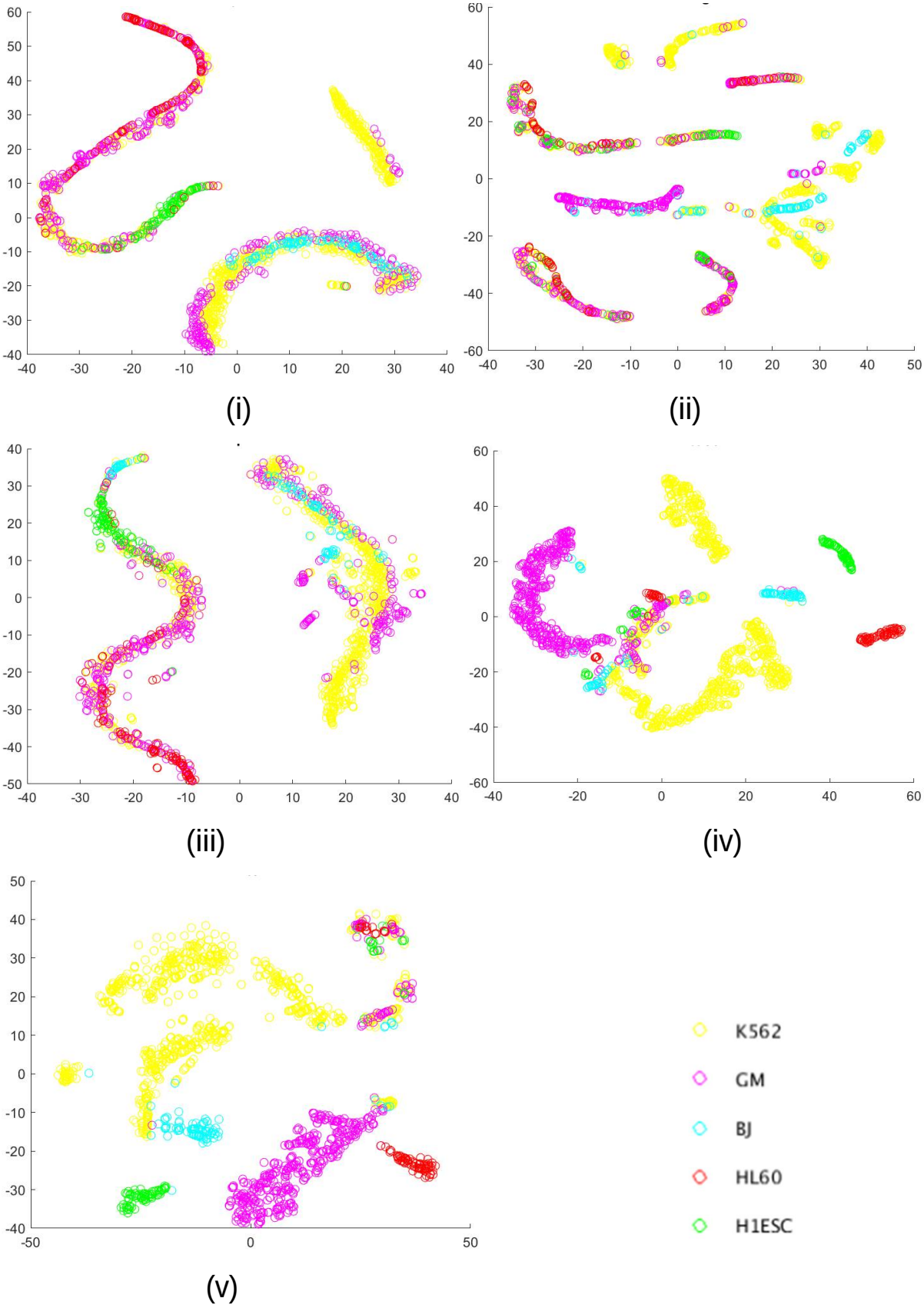


Figure 7: Visualization of single cell epigenome data for 5 cell types before and after imputation. (i) Row Data, (ii) MAGIC, (iii) scImputed, (iv) Base, (v) RITS

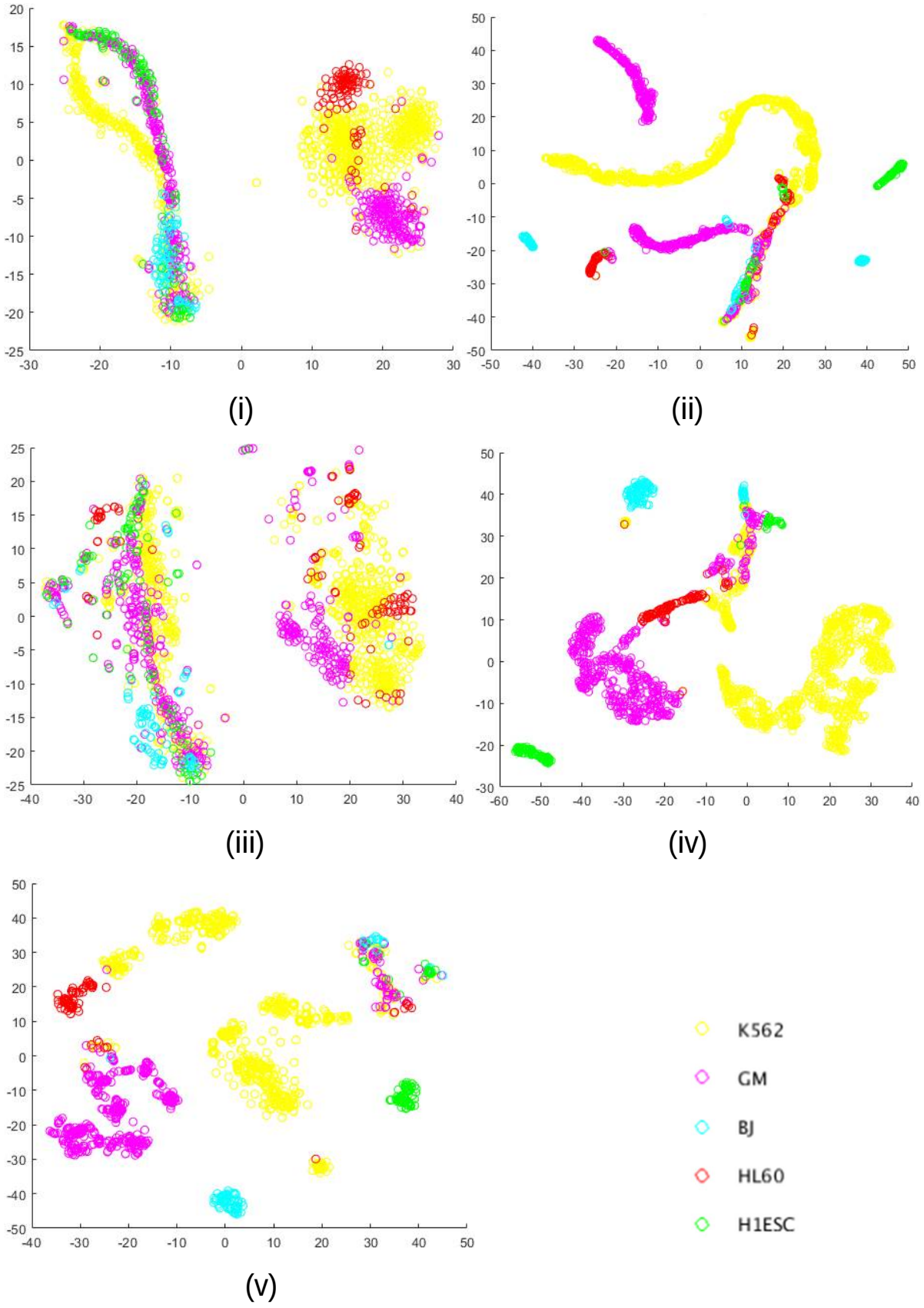


Figure 8: Visualization of single cell epigenome data for 5 cell types before and after imputation using cosine similarity as distance function. (i) Row Data, (ii) MAGIC(ideal), (iii) scImputed(ideal), (iv) Base(ideal), (v) RITS

ideal case for imputation of any scATAC-seq, hence our proposed method is more reliable in that situation. We analysed same dataset with different distance functions i.e. cosine distance and tSNE visualisation shown in figure 8. While analysing through the cosine distance metric function RITS remain at the top but in contrast to euclidean distance measure, MAGIC visualization is better and base imputation method visualisation declined as it is not able to separate different cell types clearly, as represented in figure 8(ii) and figure 8(iv) respectively. Although this time RITS separated K562 cell type into two different groups but there is no other group of different cell type present in their separation. Moreover, it can be seen through figure 8(v) RITS classified most of the identical cell types correctly even performing imputation over highly unbalanced dataset in non-ideal condition. We performed analyzation over two different distance function to gain more confidence over RITS result.

We also performed visualisation experiment over two other datasets, both are single cell ATAC-seq read counts of adult mouse from different organs i.e. Bonemarrow and liver, published by Cusanovich et al. The tSNE Visualisation over bonemarrow and liver dataset is calculated using euclidean distance measure, shown in figure 9 and 10 respectively.

From both of these figures, we observed that scImpute and MAGIC both imputation output are classified such that similar cells are assigned into different group, if their cell count is high. The figure 9(ii) visualization of MAGIC imputed data and figure 9(iii) visualization of scImputed data on scATAC –seq of bonemarrow organ of adult mouse reveals that both these imputation methods make sub-cell of highly counted cell types as ‘Hematopoietic progenitors’ and ‘Erythroblasts’ are those cell types which have high number of samples present in data. Both these cell types are sub-sampled during clustering process before imputation wrongly, hence these methods are not worth in profiling of highly unbalanced dataset. Similarly, ‘Hepatocytes’ and ‘Endothelial’ cell types are sub clustered in scATAC-seq of adult mouse liver organ data.

Base imputation method helps in identification of minor cell types in same group, only if imputation is performed with ideal parameters. Even then, it cannot separate ‘collision’ cell type perfectly for bonemarrow dataset. Imputation using RITS not performed over ideal parameters but it can clearly separate ‘Hematopoietic progenitors’, ‘collision’, ‘monocytes’ and

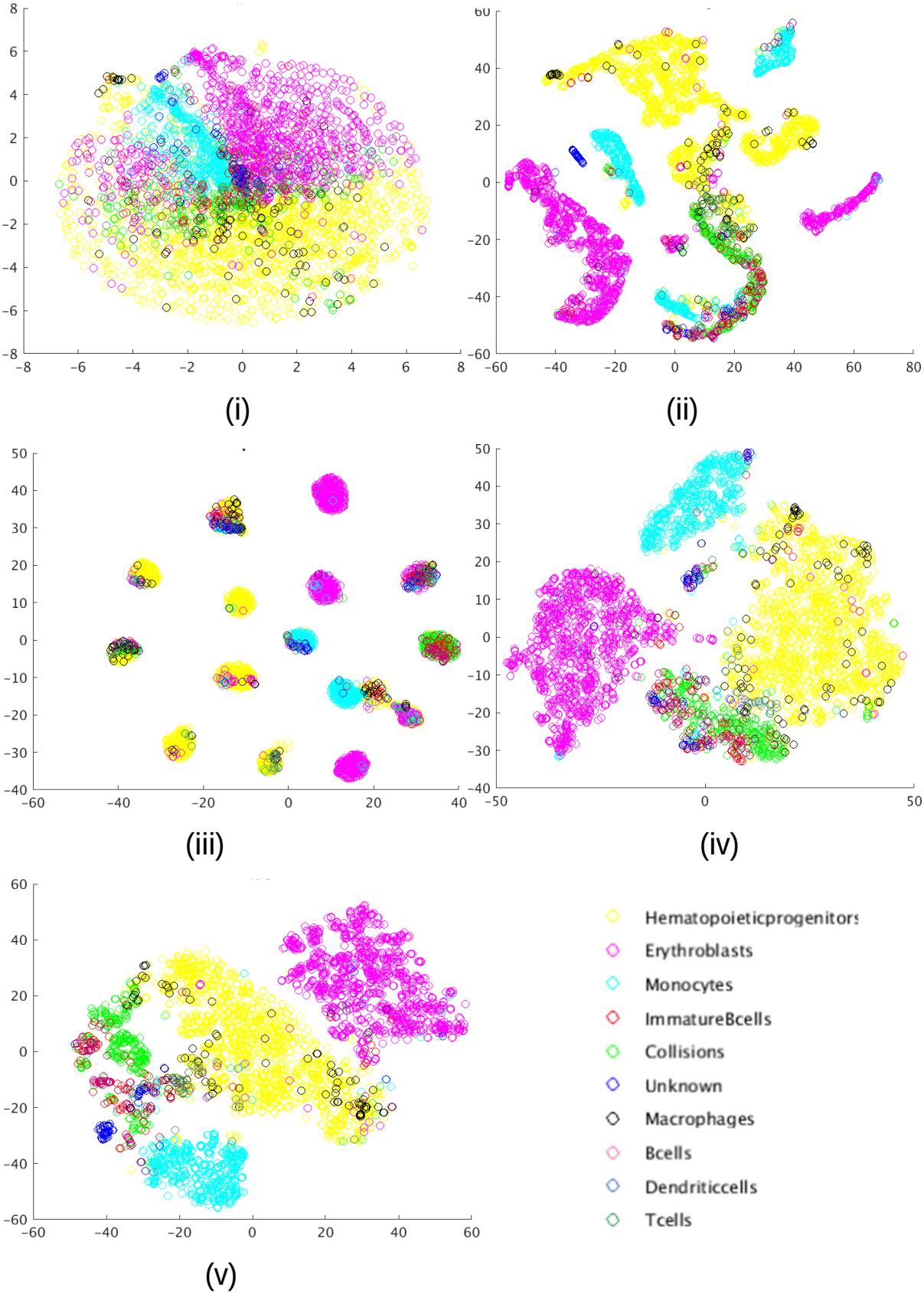


Figure 9: Visualization for ATAC-seq from BoneMarrow of adult mouse before and after imputation. (i) Row Data, (ii) MAGIC (Ideal), (iii) scImputed (Ideal), (iv) Base(Ideal), (v) RITS

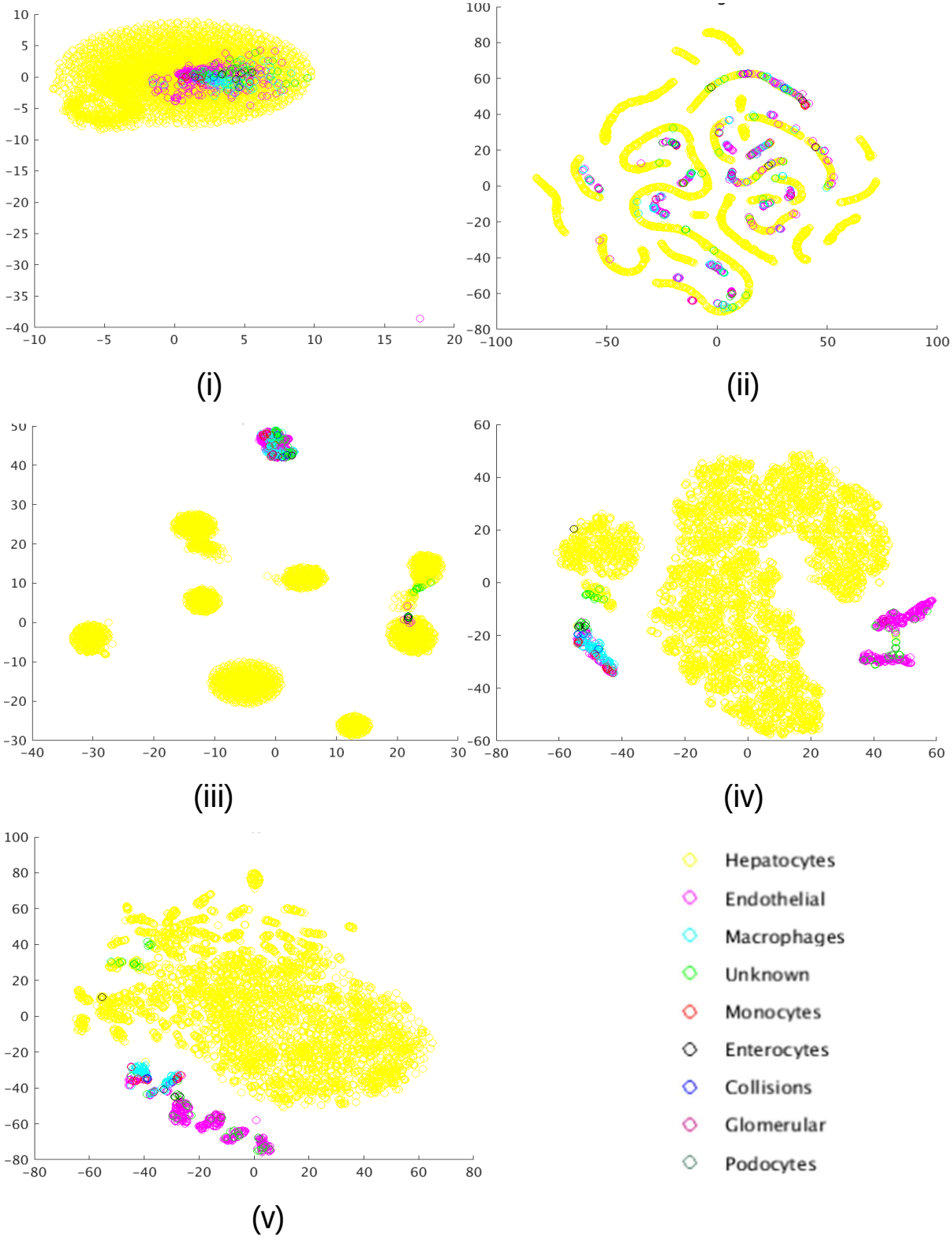


Figure 10: Visualization for ATAC-seq from liver of adult mouse before and after imputation. (i) Row Data, (ii) MAGIC(Ideal), (iii) scImputed(Ideal), (iv) Base(Ideal), (v) RITS

'Erythroblasts' cell types as compared to others. Moreover, RITS clustered 'unknown' cell types successfully.

4.5 Clustering Purity Test

Biological samples consist of different cell types, and the classification of those cell using the single cell profile is an arduous task. The efficiency of the imputation can be measured by how much it can cluster similar cells? The clustering of similar cells become more challenging for single cell epigenome data due to high noise, a high degree of sparsity and some cells belonging to different cell type but possess the same behaviour. We have used the k-mean clustering algorithm to perform clustering over non-imputed as well as imputed data. The clustering helps us to judge better, how much the imputation method is efficient in terms of enhancing inter-class differences while lowering intra-class differences during imputation. For evaluation of cluster purity, adjusted rand index (ARI) tops the list as it penalizes both false negatives as well as false positives where negative decision means two cells of similar cell type clustered into different clusters positive decision means two cells of same cell type are clustered into one cluster.

Let, U be the true partition set of 'p' cell types consisting of 'n_i' number of observations for each cell type and V be the partition obtained from result of clustering with 'k' clusters having 'n_j' number of observations in each cluster. ARI is calculated as:

$$\frac{\sum_{i=1}^{p}\sum_{j=1}^{k}\binom{nij}{2} - [\sum_{j=1}^{p}\binom{ni}{2}\sum_{j=1}^{k}\binom{nj}{2}]/\binom{n}{2}}{\left(\frac{1}{2}\right)\left[\sum_{j=1}^{p}\binom{ni}{2} + \sum_{j=1}^{k}\binom{nj}{2}\right] - [\sum_{j=1}^{p}\binom{ni}{2}\sum_{j=1}^{k}\binom{nj}{2}]/\binom{n}{2}}$$

Here, n = $\sum_{j=1}^{k} nj = \sum_{i=1}^{p} ni$ and p=k as we have used k-mean algorithm for clustering.

The figure 11 shows RARI (Relative ARI) i.e. ARI score corresponding to all methods calculated relative to maximum ARI score among them and multiplied with 100. It can be easily figured-out RITS perform excellently in this test too. This test performed over three different scATAc-seq datasets and RITS results are always at the top.

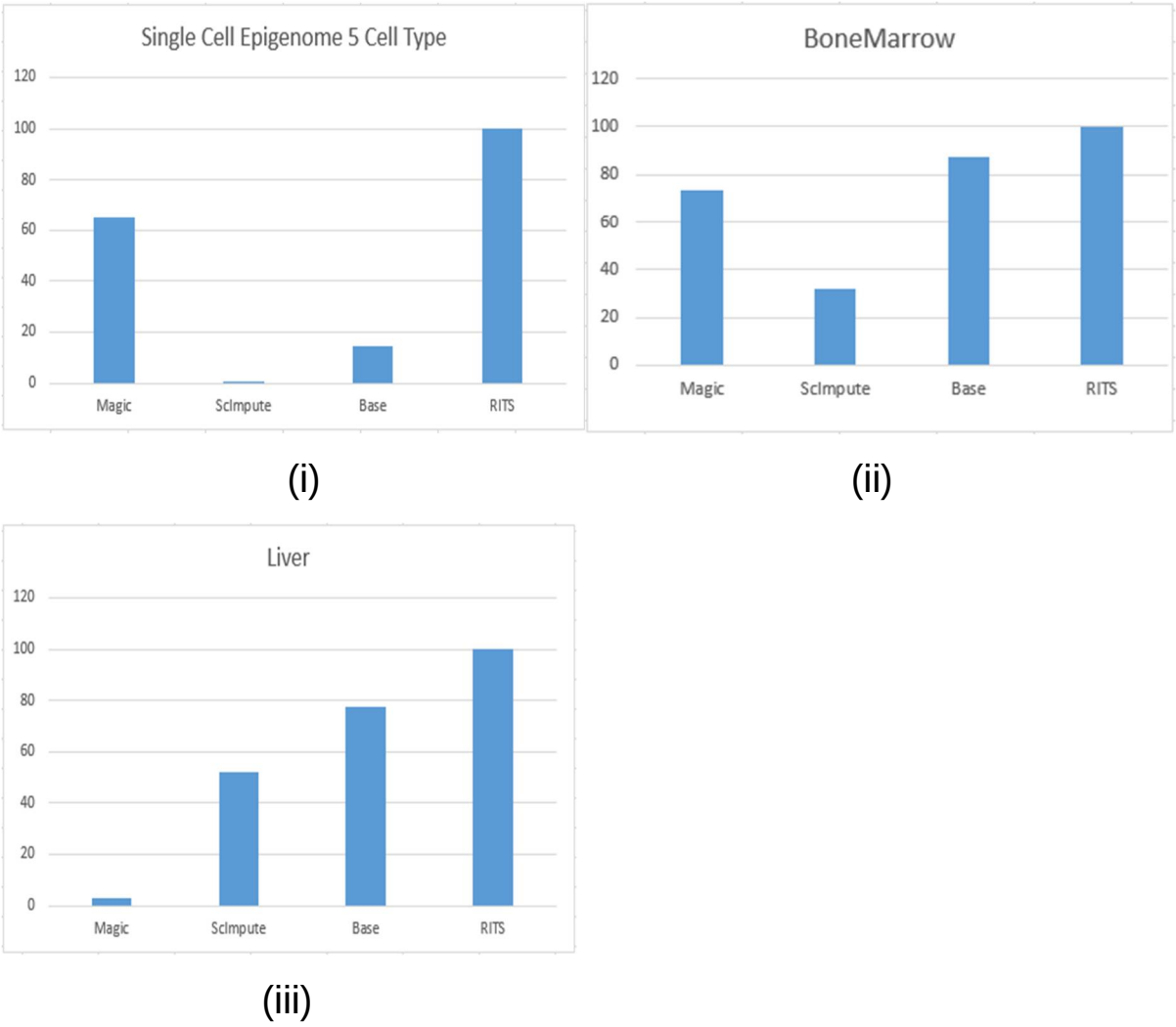


Figure 11: Relative ARI Bar Plot over different scATAC-seq datasets

4.6 Cell Type Separability Test

Similar cells have similar peak value, the better the ability of the imputation method to classify cells, the best will be its result. If any imputation method while imputing able to achieve correct clustering of cells, then it leads to an excellent prediction of missing values. Therefore, it adds one more evaluation criteria in our list of evaluation called Cell Type Separability (CTS) test. We evaluated the imputation method over inter and intra class correlation. The cells belonging to the same group have a high correlation as compared to cells belongs to a different group. CTS score is a difference between the average of the median of correlation computed for two cell groups individually called intra-cell type scatter and median of correlation

computed for pairs such that the cells belong to two different groups in each pair called inter-cell type scatter. The efficiency of the imputation method is measured by CTS score as its range lies between 0 and 1, higher its value for imputation higher the efficiency of imputation method.

We performed this test over single cell epigenome data for 5 cell types and have shown results corresponding to two pairs out of ten possible pairs. In Table 2, the test is performed between GM128 and K562 cell types as their number of cells in data are more. Although MAGIC has high score for intra class correlation for K562 but its score is also high for inter class correlation is high too. Hence for MAGIC CTS score is less than RITS, similar reason for results in Table 3.

Imputation Method	**Intra class correlation**			**Inter class correlation**	**CTS**
	GM	**K5**	**Mean**		
Row	0.057826	0.054754	0.05629	0.037077	0.019213
MAGIC	0.71914	0.78133	0.75024	0.67784	0.072395
scImputed	0.11037	0.092088	0.10123	0.059058	0.042172
Base	0.55496	0.38121	0.46809	0.20237	0.26572
RITS	0.72459	0.56519	0.64489	0.19903	0.44586

Table 2: Inter and Intra class correlation between GM128 and K562 cell type belong to single cell epigenome dataset

Imputation Method	**Intra class correlation**			**Inter class correlation**	**CTS**
	BJ	**GM**	**Mean**		
Row	0.10431	0.057826	0.081068	0.033048	0.04802
MAGIC	0.8621	0.71914	0.79062	0.70005	0.090573
scImputed	0.81749	0.11037	0.46393	0.059964	0.40397
Base	0.94981	0.55496	0.75239	0.4336	0.31879
RITS	0.8106	0.72459	0.76759	0.13612	0.63148

Table 3: Inter and Intra class correlation between BJ and GM cell type belong to single cell epigenome dataset

5. Conclusion

In this work, we have proposed an imputation method which is robust and efficient while predicting missing values in very noisy and sparse dataset. Researchers face challenge in down-stream analysis of single cell epigenome data due to the high degree of sparsity and noise. So far, there is rarely any method proposed for imputing single cell epigenome data which make RITS the first and novel tool for solving this challenging problem. Our results indicate that the proposed method outperforms other methods by a significant margin for single cell open-chromatin profiles. We have shown how our, proposed method is more reliable even while imputing highly unbalanced data. These properties make RITS the next generation imputation method for single cell genomics profiles.

Epigenome consists of chemical modification to DNA or proteins associated with DNA in a cell. The study of epigenome gained importance when researchers found that human diseases are not only caused due to bacteria or change in DNA but also by the change in the epigenome. RITS contribution makes advanced analysis of single cell epigenome data possible. Thus, RITs contribution to single cell epigenome analysis could help scientists to find differences among normal and diseased patient's cell-states.

This method can be used for imputing other types of single cell profiles as biology researchers are still facing problem while clubbing of similar cell types correctly due to high noise and similar gene expression of different cell types. We can enhance this method by using entropy and mutual information among features while performing a feature reduction. Selecting features based on high entropy and less mutual information score in combination with the random feature selection, might enhance clustering efficiency and imputation results.

Bibliography

[1]. Candes, E.J. and Y. Plan, Matrix completion with noise. Proceedings of the IEEE, 2010. 98(6): p. 925-936.

[2]. Das, S., et al., Next-generation genotype imputation service and methods. Nature genetics, 2016. 48(10): p. 1284.

[3]. Engels, J.M. and P. Diehr, Imputation of missing longitudinal data: a comparison of methods. Journal of clinical epidemiology, 2003. 56(10): p. 968-976.

[4]. Bertsimas, D., Pawlowski, C. and Zhuo, Y.D., 2017. From predictive methods to missing data imputation: an optimization approach. The Journal of Machine Learning Research, 18(1), pp.7133-7171.

[5]. Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y. and Greenleaf, W.J., 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature, 523(7561), p.486.

[6]. Clark, S.J., Lee, H.J., Smallwood, S.A., Kelsey, G. and Reik, W., 2016. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. Genome biology, 17(1), p.72.

[7]. Guo, H., Zhu, P., Wu, X., Li, X., Wen, L. and Tang, F., 2013. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome research, 23(12), pp.2126-2135.

[8]. Jin, W., Tang, Q., Wan, M., Cui, K., Zhang, Y., Ren, G., Ni, B., Sklar, J., Przytycka, T.M., Childs, R. and Levens, D., 2015. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. Nature, 528(7580), p.142.

[9]. Rivera, C.M. and B. Ren, Mapping human epigenomes. Cell, 2013. 155(1): p. 39-55.

[10]. Candès, E.J. and B. Recht, Exact matrix completion via convex optimization. Foundations of Computational mathematics, 2009. 9(6): p. 717.

[11]. Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S. and Lee, C., 2018. A single-cell atlas of in vivo mammalian chromatin accessibility. Cell, 174(5), pp.1309-1324.

[12]. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M. and Ku, M., 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 473(7345), p.43.

[13]. Gong, W., Kwak, I.Y., Pota, P., Koyano-Nakagawa, N. and Garry, D.J., 2018. DrImpute: imputing dropout events in single cell RNA sequencing data. BMC bioinformatics, 19(1), p.220.

[14]. Kumar, V., et al., Comprehensive benchmarking reveals H2BK20 acetylation as a distinctive signature of cell-state-specific enhancers and promoters. Genome Res, 2016. 26(5): p. 612-23.

[15]. Li, W.V. and Li, J.J., 2018. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nature communications, 9(1), p.997.

[16]. Majumdar, A. and Ward, R.K., 2011. Some empirical advances in matrix completion. Signal Processing, 91(5), pp.1334-1338.

[17]. Mongia, A., Sengupta, D. and Majumdar, A., 2019. McImpute: Matrix completion based imputation for single cell RNA-seq data. Frontiers in genetics, 10, p.9.

RACHESH SHARMA

MT17044, Email: rachesh17044@iiitd.ac.in
DOB: JAN 23, 1995
Address: C-6/169, Yamuna Vihar, Delhi-110053

Education

Indraprastha Institute of Information Technology, Delhi (IIITD) CGPA: 7.78
M.Tech(CSE)
2017 – Present Percentage:81.57

Bhagwan Parshuram Institute of Technology (BPIT)
B.Tech(IT)
2012 – 2016

Arvind Gupta DAV Centenary Public School, Delhi Percentage: 79.8
CBSE
2010 – 2012

DAV Public School, Delhi CGPA: 8.4
CBSE
2009 – 2010

Skills

Expertise Area	Data Structure, Machine Learning
Programming Language	Python, JAVA, Android
Tools and Technologies	NLTK, Stanford Core NLP, Android Studio, Selenium, Git, PostMan, Neo4j
Technical Electives	Machine Learning, Natural Language Processing, Pattern Recognition, Information Retrieval, Probabilistic Graphical Model

Internship

Teaching Assistant(TA),IIITD (Aug,17 – Present)
As a TA my duty is to teach students those topics which are not covered in class or most of the students have doubt. Apart from quiz, assignment and exam paper creation and its evaluation are also my task.

Software Engineer, OSSCube (july16 – may,17)
As a software engineer I developed API for Beach Body using php and mysql as backend

Projects

News Timeline Generation (Jan,18 – Apr,18)
Guide: Tanmoy Chakraborty Team Size - 4
Aim of this project was to find all the related and relevant news given a news article and creating a timeline of events for the same so that a user can precisely trace the sequence of events from the earliest to the latest news. Tools and Techniques used : Stanford

Core NLP for news preprocessing, Solr application as full text search engine, Gensim Doc2Vec for news similarity matching, Neo4j for maintaining knowledge graph.

News Timeline Generation Android App development (Jan,18 – Apr,18)
Guide: Pushpendra Singh Team Size - 5
In this project we implemented an Android app for the News Timeline generation. We used our Information Retrieval project as a service and designed the front end to make a News Timeline App.

Analyze Object Recognition Techniques (Jan,18 – Apr,18)
Guide: Richa Singh Team Size - 3
Analyze multiple aspects of objects recognition techniques and visual representation of classification accuracy through interclass separation, Like PCA , LDA , HOG Features , Neural Network Based techniques, applying Restricted Boltzmann Machine, Auto encoders and compare the result.

Quora Question Pair Similarity Detection (Aug,17 - Nov,17)
Guide: Dr. Saket Anand Team Size-3
Multiple questions with the same intent can cause problem to both seekers and writers. We use machine learning to predict whether two questions are similar.

Prediction of Reproducibility of Scientific Articles (Aug,17-Nov,17)
Guide: Dr. Tanmoy Chakraborty Team Size-2
The increase in number of non-reproducible paper is eventually creating “reproducibility crisis”. We used machine learning to curb-out this problem.

Positions of Responsibility

- Sponsor Team member of Odyssey'18 (Oct,17 – Jan,18)
  (Annual Cultural Fest of IIITD)
- Organizing Committee Member of Corona 2k16 (Jan,16 - Mar,16)
  (Annual Cultural and Technical Fest of BPIT)
- Publicity and Technical Team member of Eloquence'15 (Jan,15 - Mar,15)
  (Literary Fest of BPIT)

Interests and Hobbies

- Playing Chess
- Play Sudoku

Declaration: The above information is correct to the best of my knowledge.
RACHESH SHARMA