Published online 19 November 2020 *NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 4* **1**
doi: 10.1093/nargab/lqaa091

FITs: forest of imputation trees for recovering true signals in single-cell open chromatin profiles

Rachesh Sharma[1,†], Neetesh Pandey[2,†], Aanchal Mongia[1], Shreya Mishra[2], Angshul Majumdar[1,*] and Vibhor Kumar[2,*]

[1]Department of Electronic and Communication Engineering, Indraprastha Institute of Information Technology Delhi, Okhla Industrial Estate, Phase-III, New Delhi 110020, India and [2]Department of Computational Biology, Indraprastha Institute of Information Technology Delhi, Okhla Industrial Estate, Phase-III, New Delhi 110020, India

Received June 30, 2020; Revised October 13, 2020; Editorial Decision October 19, 2020; Accepted October 21, 2020

ABSTRACT

The advent of single-cell open-chromatin profiling technology has facilitated the analysis of heterogeneity of activity of regulatory regions at single-cell resolution. However, stochasticity and availability of low amount of relevant DNA, cause high drop-out rate and noise in single-cell open-chromatin profiles. We introduce here a robust method called as forest of imputation trees (FITs) to recover original signals from highly sparse and noisy single-cell open-chromatin profiles. FITs makes multiple imputation trees to avoid bias during the restoration of read-count matrices. It resolves the challenging issue of recovering open chromatin signals without blurring out information at genomic sites with cell-type-specific activity. Besides visualization and classification, FITs-based imputation also improved accuracy in the detection of enhancers, calculating pathway enrichment score and prediction of chromatin-interactions. FITs is generalized for wider applicability, especially for highly sparse read-count matrices. The superiority of FITs in recovering signals of minority cells also makes it highly useful for single-cell open-chromatin profile from *in vivo* samples. The software is freely available at https://reggenlab.github.io/FITs/.

INTRODUCTION

High-throughput sequencing has enabled a wider application of epigenome profiles for studying biological and clinical samples. Different kinds of epigenome profiles such as histone-modifications (1), chromatin-accessibility and DNA-methylation patterns have been used to study active, poised and repressed regulatory elements in the genome (2). Especially, for characterizing noncoding regulatory regions like enhancers, epigenome profiles have proved to be very useful (3). In the previous decade, epigenome profiling was mostly performed using bulk samples containing millions of cells. Bulk sample epigenome profiles do not help in identifying poorly characterized cell populations and rare cell types in samples of tumours or early developmental stages. Even with *in vitro* experiments, where cells differentiate, there is heterogeneity among single-cells in terms of response to external stimuli. Such heterogeneity is often not captured by using bulk epigenome profile. Moreover, heterogeneity among cells can be in both transcriptome and epigenome pattern of cells. Such as chromatin poising or bivalency at many genes may not be clearly represented through single-cell RNA-seq (scRNA-seq) profile. To explain such issues, researchers have developed techniques to profile genome-wide epigenome patterns in single-cells. Even though profiling of DNA methylation (4) and histone modification for single-cells is feasible (5), recent large scale single-cell epigenome profiles (6) have been produced using single-cell open-chromatin detection technique (7).

Single-cell open-chromatin profiling can be done using different kinds of protocols like DNase-seq (Dnase I hypersensitive sites sequencing) (8), MNase-seq (Micrococcal-nuclease-based hypersensitive sites sequencing) (9) and ATAC-seq (Transposase-Accessible Chromatin using sequencing) (10). Single-cell open-chromatin profile has the potential to reveal both active and poised regulatory sites in a genome. Most importantly, it has recently lead to an understanding of the regulatory action of transcription factors (TFs) when cells are in the state of transition (11). Besides providing a view of heterogeneity among cell states, single-cell open chromatin profiles have also proved to be useful for determining chromatin-interaction patterns (12). For analyzing single-cell open-chromatin profile, the first step is to do peak-calling after combining reads from multiple cells or using matching bulk samples. Then for each cell, the number of reads lying on the peaks is estimated. While doing so, most often researchers use a large number

[*]To whom correspondence should be addressed. Tel: +91 11 26907440; Fax: +91 11 26907405; Email: vibhor@iiitd.ac.in
Correspondence may also be addressed to Angshul Majumdar. Tel: +91 11 26907451; Fax: +91 11 26907405; Email: angshul@iiitd.ac.in
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

of peaks, sometimes exceeding more than 100000 in number (6), to capture the signal at cell-type-specific regulatory elements in heterogeneous cell-types. However, due to low sequencing depth and a small amount of genetic material from single-cells, the read-count matrix is often very sparse, which creates a demand for imputation techniques. Using a small number of hyper-active peaks to reduce sparsity may highlight only ubiquitously open sites like insulators and promoters of house-keeping genes which do not have cell-type specificity. Thus with a large number of peaks, single-cell open chromatin profiles have higher chances of including cell-type specific sites but at the cost of a high level of noise and sparsity. The sparsity in the read-count matrix of single-cell open chromatin profile is due to two reasons. The first reason is the high drop-out rate due to which many active genomic sites remain undetected (false zeros). The second reason is the genuine biological phenomenon that there is a large number of silent sites because of their cell-type specific activity. Thus, in comparison to scRNA-seq data, there are higher fractions for both true and false zeros in the read-count matrix of single-cell open chromatin profile. Given such limitations with single-cell open-chromatin profile, the classification and sub-grouping of cells is a difficult task, which is a pre-requisite for many imputation methods.

Due to the reasons mentioned above, most of the imputation methods developed for single-cell RNA-seq (scRNA-seq) profiles, could underperform on single-cell open-chromatin datasets. Hence for proper quantification of DNA accessibility using single-cell open-chromatin profiles, there is a need for a second-generation imputation method which can overcome the weakness of other such tools to handle high levels of noise and sparseness. Due to a large number of noncoding sites with cell-type-specific activity, the ideal signal-recovery method must enable detection of such sites like enhancers. Especially with recent droplet-based single-cell ATAC-seq (scATAC-seq) protocol (13), providing profiles of large number of cells with low sequencing depth, the problem of imputation becomes more eminent and challenging.

Even though there has been less attention on imputing scATAC-seq profiles, it is worth noting that many imputation methods have been proposed for scRNA-seq datasets. MAGIC (14) is the first available method for imputing scRNA-seq profiles. MAGIC predicts missing expression values by sharing information across similar cells, using the approach of heat diffusion. The approach of MAGIC involves creating a Markov transition matrix, constructed by normalizing the similarity scores among single-cells (14). While imputation of expression for a single-cell, the weights for other cells is determined using the transition matrix. MAGIC uses K nearest neighbor (KNN) approach for imputing, however unlike classical KNN-based imputation methods (15) it uses a variable value for K. Methods like MAGIC may introduce artefacts into the data and blur out genuine biological variation due to their approach of considering all zero counts as missing values. Another imputation method called scImpute (16) also tries to perform imputation on drop-out genes. For this, scImpute first learns the probability of drop-out for every gene in each cell based on a mixture model for the distribution of read-counts. scImpute predicts the missing values at false zeros by using the information of the same gene in other similar cells which it finds using genes with non-zero expression. Methods like scImpute, which use parametric method to estimate drop-out rate may not be successful for scATAC-seq in estimating true parameters due to inconsistencies in the distribution of tag-counts with very high drop-out rate and noise. High level of sparsity and noise in scATAC-seq profile reduces the chance of finding the correct neighborhood and sub-clusters of cells, which is an important step for most of the imputation methods like scImpute and MAGIC. Most recently, an approach based on deep-learning called Deep Count Autoencoder (DCA) (17) has been proposed for denoising and imputing single-cell expression profiles. DCA uses an auto-encoder to model and predicts the distribution of the genes using a zero-inflated negative binomial prior. For DCA, the mean parameter of the distribution-represents denoised reconstruction. Application of DCA on single-cell open-chromatin profile seems to be a sensible approach. However, single-cell chromatin profiles are much more sparse than single-cell RNA-seq data, hence modeling the distribution of read-counts might not always be successful.

Recently, a few methods have been proposed for visualization and clustering of scATAC-seq profiles (18). A method called SCALE (19) also uses auto-encoder to recover missing read-count values in scATAC-seq profile, whereas another tool scOpen rely on positive-unlabeled (PU) learning approach for imputation (20). Similarly, SCATE performs signal extraction and enhancement (21), but it uses previously known information such as known peaks in published bulk open chromatin profiles. Every method has its own strength and weakness such as auto-encoder-based learning is often influenced by the majority group; hence there is a chance of losing information of minor cell-types. Hence recovering missing signals in the scATAC-seq profiles is still an open problem which needs to be tackled for multiple applications of scATAC-seq in addition to visualization and clustering.

For single-cell open chromatin profiles, we realized the limitations due to improper classification and modeling of the distribution of tag-counts to estimate the drop-out rate. Therefore, we developed a method which can overcome these limitations by avoiding suboptimal solutions, using an ensemble of imputing trees. We call our ensemble-based approach as Forest of imputation trees (FITs). We have benchmarked FITs using scATAC-seq profiles of several cell types using criteria which are useful for analysis for single-cell open-chromatin. Using the scATAC-seq profile of 5 cell types, first, we show that FITs correctly recovers the chromatin accessibility of sites like enhancers with cell-type-specific activity, without performing over-imputation. We also show that FITs is more efficient than other methods in improving dimension reduction and clustering purity for scATAC-seq profiles. Further, we show that unlike other imputation methods, FITs can handle unbalanced scATAC-seq datasets and helps to avoid detecting false heterogeneity and improves detection of minor cell type. Next, we show that FITs-based restoration of the read-count matrix also helps in improving prediction in chromatin interaction using scATAC-seq profile.

MATERIALS AND METHODS

Pre-processing of data

We first check the quality of data and remove the peaks which do not have non-zero read-count in any cell. We normalize scATAC-seq read-counts and take log transform of data. Hence the read-count x_{ij} on a site g_j in cell i is represented as:

$$\bar{x}_{ij} \;=\; \log\left(x_{ij}\,/\mu_i + 1.01\right) \qquad (1)$$

where μ_i is mean read-count in cell i. The log of the normalized matrix of the read-count is provided as input to FITs. Here we have used pseudocount of 1.01 instead of 1 just like scImpute (16) to avoid the possibility of infinite values during optimization.

Clustering using randomized features in a hierarchical manner to improve imputation

Given the noise and sparsity in single-cell open-chromatin profiles finding correct subclasses is not a trivial task. Thus, we use a semi-randomized approach of clustering hand in hand with imputing. Our method has two phases: The first phase consists of making multiple imputed versions of the raw matrix through repeated clustering in hierarchical tree fashion and imputation at each node. Unlike other suggested methods, we do not use all the features at a time to perform clustering as well as we do not perform classification using the raw read-count matrix. We use an iterative approach in every tree such at every parent node we do preliminary imputation followed by classification of cells. The classification is not done only at the bottom nodes (at third layer here).

In the second phase, a final imputed matrix is assembled using the outputs from multiple trees in the first phase.

Phase-1: The first phase is described below

Given the transpose of a read-count matrix X such that cells are represented by columns and peaks by rows, we use the following approach to perform imputation in a tree:

Step1: Perform a preliminary imputation using a base method, over matrix X, taking all cells in one class.

Step2: Select n sites(peaks) randomly and perform dimension reduction using t-SNE or singular value decomposition on imputed data. Here, the number of selected peaks n is randomly chosen between 50–100% of all peaks. After reducing the dimension, apply k-mean clustering to divide cells among classes. The number of cluster k is randomly chosen in the range of two to eight.

Step3: After finding classes using k-means clustering, the raw read-count of cells in each class are assembled. After assembling the raw read-count matrix for a class, some peaks appear to have zero read-count (minimum) in all cells of that class. Hence for imputation on the raw read-count matrix of cells of a class, the peaks with all zero signals are considered as true zeros and dropped.

Step4: Imputation using base method is performed separately for cells that belong to different classes.

Step5: The imputed matrix of every class is used further to find sub-classes using the approach mentioned above in step2 and step3. Again, we randomly choose peaks (features) and value for k for the k-mean clustering.

Step6. The non-imputed raw read-count vectors of cells belonging to a subclass are assembled together in a separate matrix. Once again, the peaks which have zero read-count in all cells of a sub-class are dropped and imputation is performed separately for a matrix of each sub-class.

Step 7: The imputed read-count matrix from each sub-class is collected, and a full matrix is built. While doing so, the sites dropped in cells belonging to a subclass are given the value zero. Notice that a version of full matrix is also made using imputation for different classes at first level. Thus from every tree, we collect two versions of the imputed matrix.

The above steps 1–7 are repeated many times to get an ensemble of imputation trees.

The output from several trees from phase-1 is further processed in phase-2 using the steps described below.

Phase-2: Following steps are taken in phase 2

For every cell correlations between its unimputed read-count vector and imputed versions from Phase-1 are computed. For every cell average of m most correlated imputed versions, is taken as the final imputed vector. For $m = 1$, one has just to take the topmost correlated imputed version. The user decides the value of m, and it can range from 1 to the number of trees made in phase-1. Here we have used the default value of $m = 3$ for benchmarking FITs on different datasets. If the number of imputed version is less than 3, and the user does not provide an option ($m = 1$), it uses all of them to make the final vector.

The reason and logic for some steps in phase-1 and phase-2 are explained below:

i. At every node of the tree, initial imputation is done before dimension reduction and clustering, so that chance of getting the correct cluster is high.

ii. Further, sub-classification is done so that cells belong to a minor cell-type or cell-state could get chance to come together to have more accurate imputation.

iii. The imputed version of read-count at level-1 of a tree is also collected so that if a cell belongs to a majority class, we should not force its imputation using smaller groups of cells.

iv. In phase-2, we use spearman correlation to choose best k imputed version. We tried several kinds of distance measure to calculate the similarity between unimputed and imputed read-count vectors and found that spearman correlation-based selection of the most suitable imputed version provided the best results.

The step of choosing imputed vectors, which have the highest correlation with unimputed read-count, is inspired by the minimization criteria followed by nearly all imputation methods. The classical imputation methods based on finding lower rank matrix, the difference between imputed and non-imputed matrices is minimized at observed features, to avoid under and over-imputation. Thus, in other words, we can say that FITs applies the minimization criteria two times, one during imputation at every node of trees in phase-1 and other at the stage of phase-2.

The base imputation method of FITS

Even though FITs is designed to be robust to handle error caused during imputation, it is worth describing the underlying base imputation method used by FITs. The base imputation method uses the approach of nuclear norm minimization with singular value soft-thresholding, as explained below.

Given a read-count matrix Y of a set of cells, where columns represent peaks and rows are for individual cells. The observed read-count matrix Y can be called a sampled version of true ideal matrix X. It can be represented as:

$$Y = A(X) \tag{2}$$

Here A is an operator matrix which causes sub-sampling, and has 0's where the elements of X is not observed, and 1's where it is known. The problem of imputation here is to recover complete matrix X, given the read-counts in Y, and the sub-sampling mask A.

Most often approximate rank of the matrix X is not known, so getting a solution for equation (2) is not easy. In order to resolve these issues, researchers use an alternative solution. For this purpose, researchers try to solve equation (2) with a constraint that the solution is of low-rank. This mathematical representation for this can be written as,

$$\text{minrank}\,(\mathrm{X})\,\text{such that } Y = A(X) \tag{3}$$

However, this problem itself is NP-Hard. Therefore its closest convex surrogate; nuclear norm minimization is used by many studies (22,23) for matrix completion. The nuclear norm minimization can be termed as:

$$\min_{X} ||X||_{*} \text{ such that } Y = A(X) \tag{4}$$

Here $||.||_{*}$ represents the sum of singular values of data matrix X and is called as nuclear-norm. This constraint of minimum rank can be replaced with l_1 norm of the vector of singular values of X as a stringent and convex alternative. Hence as a solution, a modified version of the above equation is proposed (23) as:

$$\min_{X} ||\, Y - A(X)\,||_F^2 \; + \lambda||X||_{*} \tag{5}$$

Here λ is the Lagrange multiplier. There is no closed-form solution for the problem in equation (5). Therefore, it is solved in many iterations. To solve such problem, we use majorization–minimization approach at iteration k, given below

$$\min_{X} ||\, B - X||_F^2 \; + \lambda||X||_{*} \tag{6}$$

$$\text{Where } B_{K+1} = X_K + \frac{1}{a} A^T (Y - A(X_K)) \tag{7}$$

Using the inequality $\min_{X} ||\, M1 - M2|| \rangle ||s1 - s2||$ where s1 and s2 are singular values of matrices M1 and M2, we can express the minimization problem as

$$\min_{X} ||\, s_B - s_X||_2^2 \; + \lambda||s_X||_{*} \tag{8}$$

Where s_B and s_X represent singular values of B and X, respectively, and $||s_X||$ is the sum of absolute of singular values of X (24). Thus the minimization problem in equation (7), is often solved by soft thresholding (24) in the following manner

$$\mathrm{s_X} = \mathrm{sign}\,(\mathrm{s_B})\;\max\,(0, |\mathrm{s_B}| - \lambda/2) \tag{9}$$

It has been found that the algorithm is robust to the value of λ as long as it is reasonably small (25).

Estimating co-accessibility among sites and evaluating the prediction of chromatin interaction

Given a read-count matrix of single-cell open-chromatin profile, if we have to find co-accessibility among genomic sites, we can calculate a covariance matrix. However, as the number of elements to be calculated in the covariance matrix is usually larger than the number of data-points in the read-count matrix, estimating the true covariance matrix is not a trivial task. Moreover, the covariance matrix may not always represent direct interaction among genomic sites. Therefore, Graphical Lasso (26) is quite suitable for this kind of problem. The Graphical Lasso method helps in estimating regularized covariance and inverse of the covariance matrix, which can be used to calculate partial correlations between variables (genomic site) (26). The partial correlation represents the measure of the degree of association between two variables when the effect of other variables is removed. Given the noise and small size of data, Graphical Lasso aims to detect a small fraction of true partial correlations among variables. It uses a penalty term which cause shrinkage of partial correlations between many pairs to value zero, if there is not enough strength in the estimate of their association. Graphical Lasso aims to minimize:

$$\text{logdet}\Theta - tr\,(U\Theta) - \rho||\,\Theta||_{\,1} \tag{10}$$

Where Θ is the inverse covariance matrix having the dependence structure of variables and U is their covariance matrix, and ρ is the penalty term for L1 norm-based regularization. Unlike Cicero (by Pliner *et al.* (12) we did not use the technique of having a penalty term dependent on the distance between genomic sites, as we did not want to miss distal interaction. Moreover, our target here was just to evaluate the improvement in the prediction of co-accessibility by imputation. Here we used the value $\rho = 0.01$.

Before estimating co-accessibility, we merged peaks which were within 25 kbp of each other and also added their read-counts. In other words, read-counts in bins of 25 kbp were used. We performed this task on both imputed and non-imputed read-count matrix before calculating their covariance matrix and applying Graphical Lasso. We calculated partial correlation values (co-accessibility scores) between each pair of the genomic region (merged peaks or bins), using the inverse covariance matrix estimated by Graphical Lasso. We downloaded the processed chromatin interaction files for K562 and GM12878 provided by Rao *et al.* (27) in HiC data format (.hic format). Using files in .hic format, we derived the interaction using Juicer (28) and converted the output to six column bed format with scores. Out of all interactions, we chose high confidence interactions with *P*-value < 1E-9. We used PGLtool (29) to find overlap between high-confidence HiC-based chromatin interaction and predicted interacting peak-pairs using co-accessibility.

Evaluation measures for separability and clustering

After t-SNE (t-distributed Stochastic Neighbor Embedding) (30)-based dimension reduction of the imputed and non-imputed read-count matrix, we performed k-means clustering. We used two measures to judge the different properties of clustering and imputation. The first method called adjusted Rand index (ARI) has cost for false positive and false negatives, where 'positive' means that cells of the same type are clustered into one cluster and 'negative' means that two similar cells are assigned different clusters. Let, $T = [t_1, \ldots, t_P]$ represents the true p classes consisting of n_i number of observations in class t_i and $V = [v_1, \ldots, v_K]$ be the clustering result with 'k' clusters having n_j number of observations in cluster v_j. ARI is calculated as:

$$\frac{\sum_{i=1}^{p}\sum_{j=1}^{k}\binom{n_{ij}}{2} - \left[\sum_{j=1}^{p}\binom{n_i}{2}\sum_{j=1}^{k}\binom{n_j}{2}\right]/\binom{n}{2}}{\left(\frac{1}{2}\right)\left[\sum_{j=1}^{p}\binom{n_i}{2} + \sum_{j=1}^{k}\binom{n_j}{2}\right] - \left[\sum_{j=1}^{p}\binom{n_i}{2}\sum_{j=1}^{k}\binom{n_j}{2}\right]/\binom{n}{2}} \quad (11)$$

Here, $n = \sum_{j=1}^{k} n_j = \sum_{i=1}^{p} n_i$

The second measure we used is called as cell type separability (CTS). To calculate CTS, first, we find spearman-correlation between read-counts of each possible cell pair. Then we calculate the median correlation for pairs of cells belonging to the same type to get the intra-cell-type median correlation. Then among two cell-types, we calculate inter-cell-type median correlation by taking only those pairs where one cell belong to one of the two types. The difference between intra-cell-type and inter-cell-type median correlations is called as CTS. Thus, CTS value is always calculated between two cell-types.

Data description and availability

The datasets used in this manuscript are available in public repositories. The reads for a scATAC-seq profile published by Buenrostro *et al*. (10) were downloaded in SRA format (SRA ID: SRP052977). The reads were aligned to hg19 version of the human genome using Bowtie (31). Peaks of ATAC-seq human cell types used here are provided in the same study by Buenrosto *et al*. at GEO database (GSE65360). The peaks were merged using bedtools. The read-count for every cell was then estimated for peaks in the merged peak list. The scATAC-seq read-count for immune cells was downloaded from GEO database (ID: GSE96772) (32). Single-cell ATAC-seq read-counts for cells from Bone-marrow and liver of adult mouse is available with GEO ID: GSE111586. For the dataset of same GEO id: GSE111586 we also used pathway enrichment-score-based analysis. It had annotations for approximately 39 major cell-types including 'collision' and 'Unknown'.

For evaluation, we used peak-list of bulk sample ATAC-seq profile of three cell types BJ, GM12878, H1ESC from other GEO database (GEO ID: GSE65360). Other bulk ATAC-seq profile used here had GEO IDs as such BJ: GSE113414, GM12878:GSM1155958, H1ESC:GSM2083754, HL60:GSM2083754 and K562: GSM1782764. For defining enhancers, ChIP-seq peaks of histone modification H3K27ac were used. The H3K27ac peak-list for H1ESC, GM12878 are made available by ENCODE consortium and made available in UCSC genome browser (1). The peak-list for H3K27ac ChIPseq for HL60 is available with GEO IDs: GSM2418804 (33). The chromatin interaction files for K562 and GM12878 cell lines downloaded in .hic format have been made available by Rao *et al*. (27) (GEO ID: GSE63525) (27).

RESULTS

Biologically similar cells would have similar activity level at a regulatory site, and this fact can be used to impute the missing values. Hence, if we group similar cells in a sub-cluster, an imputation method has a high probability of providing correct results. However, given the noise, sparsity and imbalance in single-cell open-chromatin dataset, achieving correct sub-cluster is not a trivial task. Hence our method uses randomization with multiple hierarchical tree-based clustering hand-in-hand with imputations using a base method (Figure 1). The base imputation method used at every node in the tree uses a known procedure of soft thresholding of singular values for matrix completion (see 'Materials and Methods' section). The hierarchical tree-based approach used by our method is such that we first perform an initial imputation for all the cells taking them in one group. Using the initial imputed read-count matrix, we classify the cells in into K classes (nodes). For cells belonging to each class, we perform imputation using their raw read count and ignore the previously imputed matrix. However, when we assemble the read count of cells belonging to a particular class, multiple peaks (genomic sites) have zeros read count in all cells belonging to that class. This is exactly as expected, and we utilize it to improve imputation. We consider those peaks with zero read counts in all cells in a class as true-zeros and drop them while doing class-wise imputation. Again, we use the imputed read-count of non-dropped peaks of cells belonging to a class for further classification. Thus, we get subclasses of cells and we again group the raw read count of cells belonging to a subclass. One important point to be noted is that for every level of classification, we randomly choose 50–100% of the non-dropped peaks for classification. We also randomly choose the number of classes k in k-mean clustering at every step. Thus, we perform many such hierarchical tree-based clustering and imputation while randomly deciding k (number of classes) and features for classification. After having final imputed matrices from many such trees, we use the best jth column of multiple imputed version of read-count matrix, for a cell j based on correlation with raw read-count. It is based on our observation that spearman correlation between unimputed read-count vectors of cells of the same type is higher than the correlation between non-similar cells (see Supplementary Figure S1). Hence, for a cell, if the imputation is done by clubbing it with wrong neighbors, the imputed vector will have a lower correlation with the unimputed version in comparison to correct imputation. The last step of choosing the best *m* vectors from multiple imputed version is a crucial filtering step which further helps FITs in avoiding over-imputation (Figure 1). The motivation and logic of

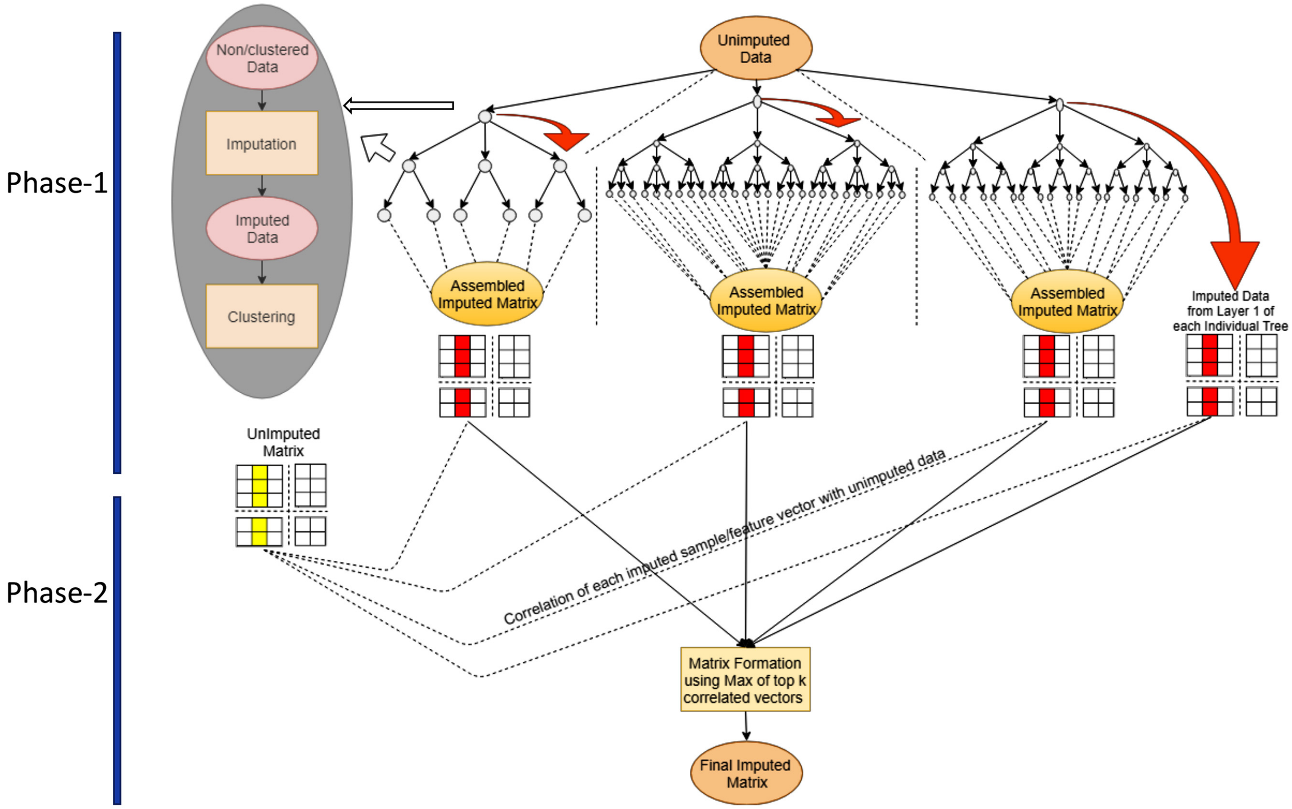

Figure 1. A description of FITs: FITs has two phases. In phase-1 many imputation trees are built to get different imputed versions of the read-count matrix. In an imputation tree, at every node, first, an imputation is performed on the non-imputed read-count matrix using a base method, followed by dimension reduction. Then *k*-mean clustering is performed to get *k* clusters of cells. The raw-read count of cells of each cluster is passed on to one daughter node where the same procedure of imputation and further clustering is followed. At every node of an imputation tree, the sites with zeros in all the cells in the raw read-count matrix, are dropped. In phase-2 of FITs, the vectors of multiple versions of imputed matrices (shown as red column) are compared to corresponding vectors in the original unimputed matrix (shown as yellow column) using correlation. Finally, for every cell, only those imputed versions are taken which have the highest correlation with its unimputed read-count vector.

different steps of FITs are provided in detail in the 'Materials and Methods' section.

FITs recovers open chromatin signal and avoid over-imputation

We compiled read-count matrix of scATAC-seq profile published by Buenrostro *et al.* (10) (see 'Materials and Methods' section). Our compiled datasets had five cell types (GM12878, K562, HL60, BJ, and H1ESC) and consisted of 1622 cells and 92 447 peaks. We first evaluated if the application of FITs, improves the data quality of single-cell ATAC-seq by correlating it with the relevant bulk ATAC-seq profile. We found that FITs-based signal-recovery increased the correlation among bulk and single-cell ATAC-seq profiles (Figure 2A). For different cell-types in the compiled dataset (GM12878, H1ESC and K562, BJ and HL60) cells, there was almost 4-fold increase in correlation between scATAC-seq and bulk ATAC-seq profiles after application of FITs. Next, we evaluated FITs using promoters of markers genes which are expected to have open-chromatin in a cell-type-specific manner. FITs was able to improve the read-count signal of cell-type-specific promoters without over-imputation in other cell-type. Such as for promoter of CD79a, which has B-cell-specific expression (34), FITs caused amplification of its read-count signal only in GM12878 (Figure 2B). Similarly, for the promoter of the SOX2 gene, the imputation by FITs caused an increase in read-count value only for H1ESC (Figure 2C).

We further evaluated the performance of FITs in comparison to KNNimpute, and three other methods (MAGIC, scImpute and DCA) developed for single-cell RNA-seq read-count matrices. For every genomic site in the used scATAC-seq dataset, we first found in which of the five cell types it overlapped with a genuine peak of bulk ATAC-seq profile. We estimated the coverage for peaks of bulk ATACs-seq in respective cell-types and calculated ROC-AUC for every cell (Figure 3A). FITs-based signal-recovery resulted in consistently higher median AUC than other imputation methods for coverage for true peaks from bulk samples.

FITs improves detection of cell-type-specific sites

One of the main purposes of open chromatin profiling is to study the activity of cell-type-specific regulatory elements like enhancers (35). Few scientific groups have used the technique of highlighting cell-type-specific activity using open chromatin signal to predict enhancers (8). We used a similar technique and divided scATAC-seq read-counts on a peak by its average read-count across all the cells. For vali-

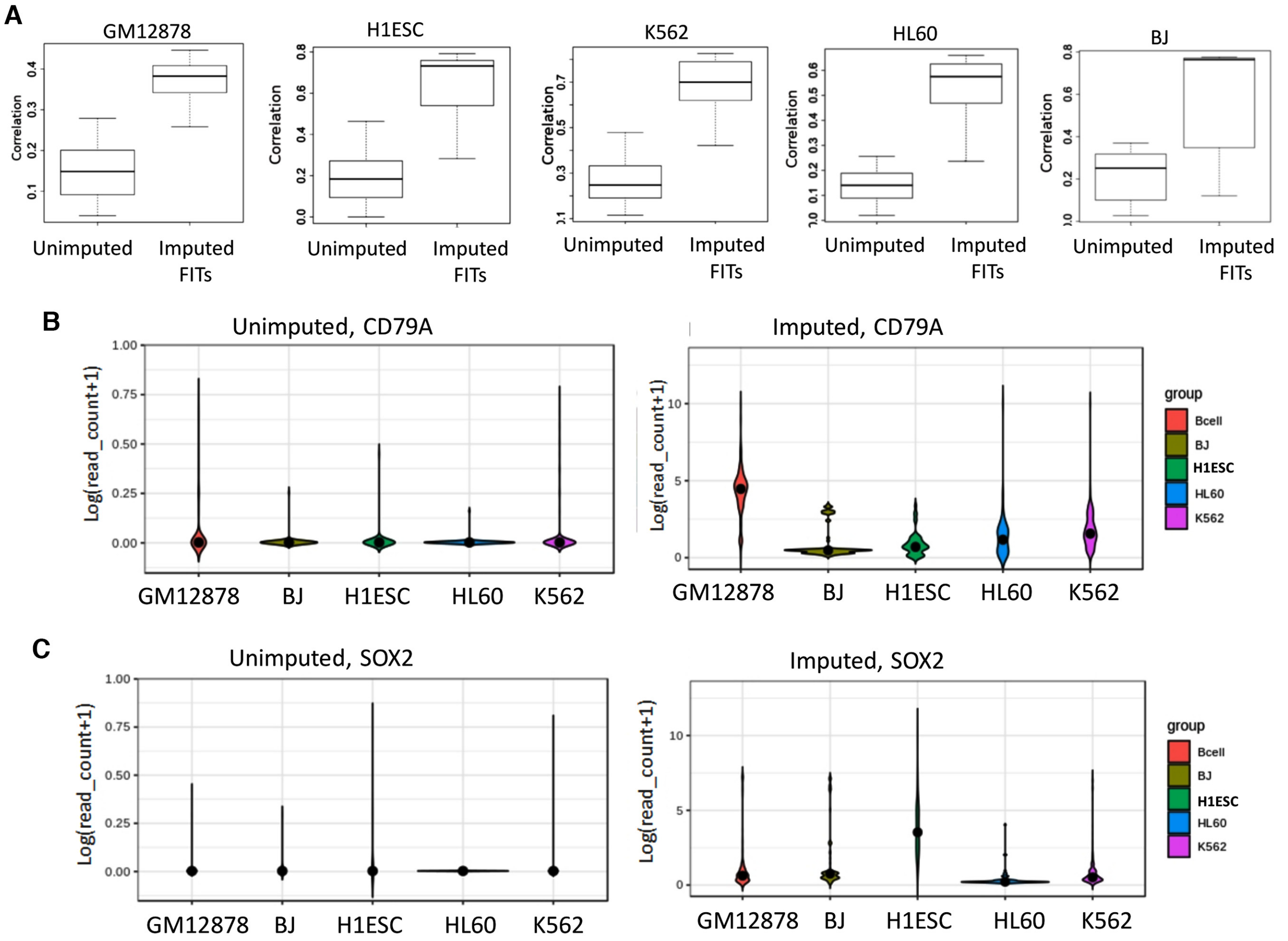

Figure 2. FITs improves signal in single-cell ATAC-seq profile. (**A**) Boxplot of correlations of imputed and non-imputed scATAC-seq read-count with bulk ATAC-seq profile in relevant cell-types from other studies. Results for five cell types, GM12878, H1ESC, K562, HL60 and BJ are shown here. (**B**) Violin plot of imputed and non-imputed read-count on the promoter of CD79a which is known to have expression more specifically in B-cells (GM12878 here). (**C**) Violin plot of read-counts on the promoter of SOX2 gene. Among five cell types in our dataset, SOX2 gene is supposed to have expression only in H1ESC. FITs-based imputation shows high read-count of SOX2 only in H1ESC cells.

dation, we used non-promoter peaks of H3K27ac ChIP-seq profiles of bulk samples of cell-lines, as enhancers (1,33). Further evaluation and comparison with the other four methods revealed that imputation with FITs consistently provided higher coverage for enhancers compared to other methods (Figure 3B). DCA had a comparable performance for H1ESC cells; however, for BJ cells, DCA seems to have failed in recovering a genuine signal (Figure 3B). The performance of MAGIC varied for different cell types, whereas median AUC for scImpute for detection of enhancers remained low in the range of 0.52–0.62 (Figure 3B).

Overall FITs restores signal at cell-type-specific sites in scATAC-seq profiles without over-imputing, which can help researchers to detect enhancers for downstream analysis. The capacity of improving signal at cell-type-specific sites can help in improving CTS. Here, CTS is defined as the difference between the median intra-cell-type correlation and inter-cell-type correlation (see 'Materials and Methods' section). We calculated CTS score among different pairs of cell types (BJ versus GM12878; BJ versus H1ESC and GM12878 versus H1ESC) and found that FITs-based imputation provided the best CTS among all four methods used for comparison (Figure 3C). Among other methods, DCA appeared to be second best, but the CTS values for DCA were substantially lower than FITs.

FITs improves dimension reduction and clustering of single-cell ATAC-seq profiles

One of the major tasks in the analysis of single-cell open-chromatin profile is to reduce the dimension of read-count matrix for visualization and classification. Due to the high level of noise and sparsity in scATAseq read-count matrix, researchers often resort to calculating accessibility score for TF motifs for dimension reduction-based visualization and classification (6). However, dimension reduction and classification of read-counts directly could reveal new classes and states of cells which could be blurred out by using motif accessibility scores. We performed t-SNE (30)-based dimension reduction and visualization of scATAC-seq read-count matrix for five cell lines published by Buenrostro *et al.* (10). As expected, the application of t-SNE on raw

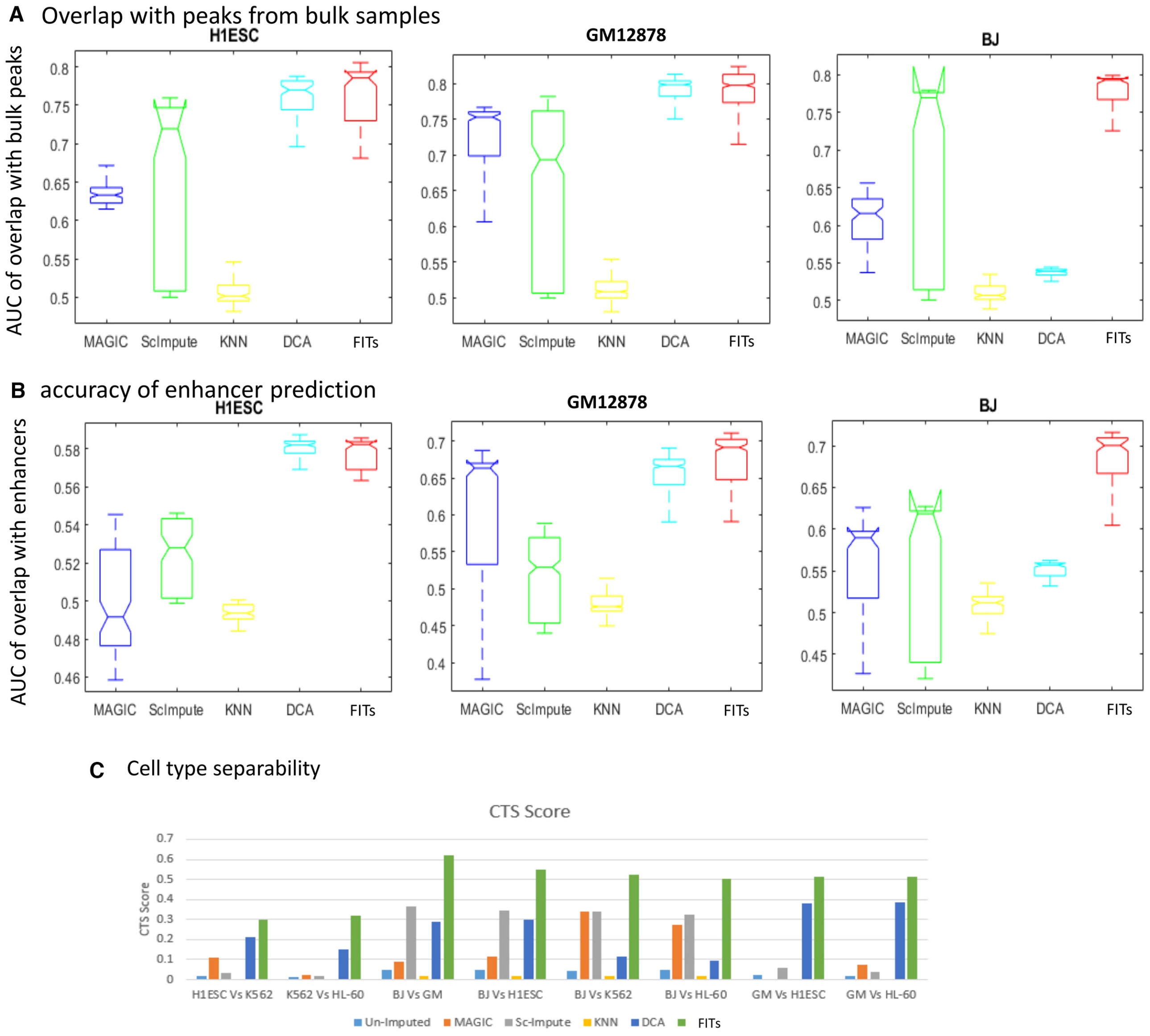

Figure 3. FITs-based imputation improves coverage of true peaks and enhancers. For evaluation single-cell ATAC-seq dataset published by Buenrostro *et al*. (10) was used here. (**A**) For every cell coverage for relevant bulk ATAC-seq peak according to intensity of read-count in the imputed read-count matrix was calculated using the approach of ROC (receiver operating characteristic curve). Here positive value means that a site has a peak and negative represent no-peak in relevant bulk ATAC-seq profile of the relevant cell. The area under the ROC (AUC) for each single-cell was calculated and box plots of AUC of cells for different cell types are shown here. (**B**) Boxplot of AUC for coverage of enhancers using normalized read-count of scATAC-seq profile. The true set of enhancers for a cell-type was compiled using H3K27ac ChIP-seq profile from the bulk sample. (**C**) CTS among cells of different cell-types calculated using imputed read-count matrices.

(unimputed) read-count did not provide satisfactory results as cells of different types were co-localized together in lower-dimensional space (Figure 4A). Similarly, applying t-SNE on read-count matrix imputed by other tools methods also provided results which had a mixing of coordinates for cells of different types. However, with read-count matrix imputed by FITs, the coordinates provided by t-SNE had clear separability among different cell types. It is quite evident from t-SNE plots of imputed matrixes that MAGIC and scImpute introduce artifactual grouping of cells (Figure 4A). MAGIC and scImpute outputs had artefacts possibly due to complete reliability on one-time grouping and sub-classification of the raw read-count matrix for imputation. In the output based on KNNimpute, GM12878 and K562 cells appeared to have overlapping locations in t-SNE-based visualization. On the other hand, DCA seems to have mixed the profile of H1ESC and GM12878 during imputation (Figure 4A). Further, we compared the accuracy of clustering using the imputed scATAC-seq profiles. For this purpose, we used the ARI after applying k-means clustering on t-SNE-based coordinates for imputed read-count matrices. FITs had highest ARI score among the tested methods. The ARI scores for output of other methods were two to three times lower than FITs. When *k*-means clustering was used after spectral embedding (36), to get the same number of clusters, the results were similar and FITS had better ARI

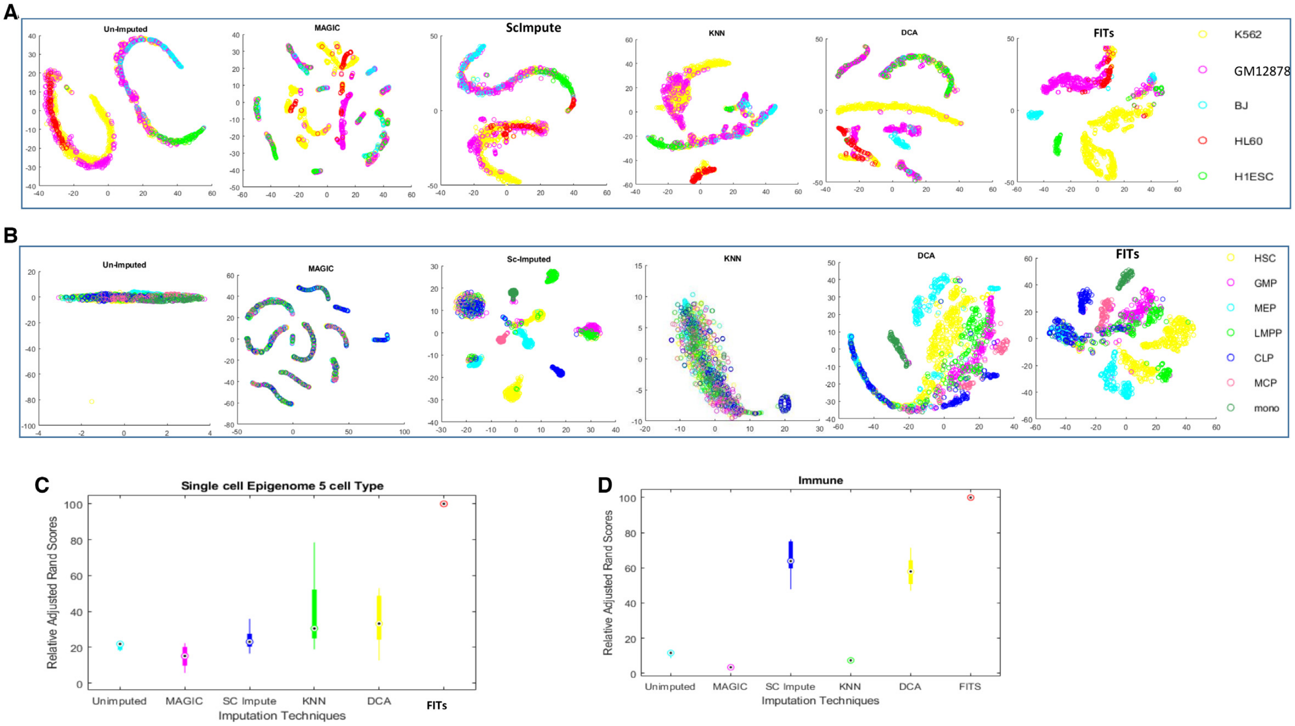

Figure 4. Performance of different imputation methods based on t-SNE-based dimension reduction. (**A**) Scatter plot of result of t-SNE for outputs from different imputation methods for single-cell ATAC-seq profile from Buenrostro *et al.* (10) (**B**) Visualization of t-SNE results for unimputed and imputed read-count matrices for hematopoietic cell dataset. (**C**) Boxplot of Relative ARI to enumerate clustering quality when *k*-means clustering is applied on t-SNE-based reduced dimension. The results clustering of ATAC-seq profile from Buenrostro *et al.* (10) is shown here. To make boxplot, ARI was calculated at different values of k in the range of 3 to 10. The absolute values of ARI at $k = 5$ are; Unimputed: 0.066674, MAGIC:0.017608, scImpute:0.057308, KNNimpute:0.16264, DCA:0.10685, FITS: 0.20714. (**D**) Boxplot of ARI relative to FITs for evaluation of classification of hematopoietic cells. The relative ARI values were calculated at different values of *k* (between 5 and 12) for *k*-mean clustering. The absolute values of ARI for immune cells data at $k = 10$ are; Unimputed:0.033804, MAGIC:0.012723, scImpute:0.17961, KNNimpute:0.024815, DCA:0.19065, FITS:0.37533.

score than other imputation method-based result (Supplementary Table S1).

We simulated higher drop-out rate using the same scATAC-seq dataset of 5 cell lines (by Buenrostro *et al.* (10) and randomly dropped 10–30% of read-counts. At all simulated additional drop-out level FITs output provided clear separability among different cell-types in t-SNE-based visualization (Supplementary Figure S2A–C). With a higher drop-out rate in five cell-line datasets, the performance gap between FITs and other methods further increased in terms of clustering purity achieved after imputation. As can be seen in Supplementary Figure S2D, the ARI score for FITS output is three to four times greater than all other methods compared for simulated drop-out. Further, we used a dataset which had a severe problem of drop-out as well as overlap among the activity of regulatory sites among different cells. We also used scATAC-seq dataset generated using phenotypically-defined human hematopoietic cells, including early hematopoietic-progenitors and cells of myeloid and lymphoid lineage (32). This dataset of immune cells had cell-type labels for every cell. After applying different imputation methods followed by t-SNE-based visualization, we achieved similar results such that FITs had more separability among non-similar cell-types (Figure 4B). Whereas, unimputed version and outputs of MAGIC and KNNimpute had completely mixed co-localization for different cell types. In terms of ARI even scImpute and DCA were not comparable to FITs (Figure 4D). ARI scores calculated after spectral-embedding-based clustering (36) also showed better performance of FITs compared to other five imputation methods (Supplementary Table S1). On careful observation, we found that there are two cell types (HSC and LMPP) which have two groups of cells in t-SNE plot for FITs (Figure 4B). Our investigation revealed that those groups came from different batches which might have different culture-micro-environment or experimental setup. Thus, the process of restoration of open-chromatin signal and reduction in noise by FITs improved clustering and separability of different cell-states and batches.

FITs can handle unbalanced read-count matrices and restore signal of minor cell population

The approach of hierarchical steps of imputation and clustering hand-in-hand with randomization in FITs has the potential to highlight minor population cluster even in the presence of dominating signal from major cell-types. In order to evaluate the performance of imputation methods on imbalanced dataset of scATAC-seq profile, we first created a dataset consisting of K562 cells with 95% frequency and H1ESC with 5% occurrence rate. After imputation of the simulated dataset, we found that FITs was able to recover the signal of most of the cells of minor cell-type (H1ESC) such that they could localize as a separate group in t-SNE-based visualization (Supplementary Figure S3). Whereas in the case of other imputation methods, H1ESC co-localized with major cell-type (K562) in t-SNE plot of imputed dataset (Supplementary Figure S3). The output of FITs also had highest ARI score for clustering purity of imputed imbalanced dataset.

After evaluating imputation methods for simulated imbalanced dataset, we used *in vivo* datasets for further testing. We performed imputation on scATAC-seq profiles of cells from adult mouse bone marrow and liver published by Cusanovich *et al.* (6). For scATAC-seq dataset of adult mice, Cusanovish *et al.* have performed annotation and assigned cell-type for most of the cells. The compiled scATAC-seq dataset for bone marrow had 4033 cell and liver data had 6167 cells. After imputation, we performed t-SNE-based dimension reduction and visualization using eight most frequent cell-types and cells with a label of 'unknown', for both datasets. Even among retained cell-types, there was an imbalance in numbers. In scATAC-seq dataset of bone marrow, there were four cell types (Macrophages, B cells, Dendritic cells and T cells) with a frequency <2% and two cell types represented more than 70% cells (Hematopoietic progenitor + erythroblast). In liver scATAC-seq data, there was an even higher imbalance in numbers for cells of different types. Among the retained cells in liver scATAC-seq data, 91.4% were hepatocytes, while 6 other cell-types had frequency <1.6%.

Visualization of t-SNE results for bone marrow datasets revealed other tested imputation methods were inefficient in recovering signal of minor cell-types (Supplementary Figure S4). For MAGIC, scImpute and KNNimpute, minor cell type locations got mixed with major cell-types in t-SNE plots. For bone marrow scATAC-seq profile, the imputation by scImpute caused the formation of many clusters within major cell types such as erythroblast, hematopoietic progenitors and monocytes. On the other hand, minor cell types in bone marrow data such as B cells, dendritic cells could not be isolated as separate groups in t-SNE results for matrix imputed by scImpute. MAGIC also had similar results like scImpute in terms of separability for minor cell types. However, for bone marrow data, FITs was efficient in recovering signal even for minor cell types. The t-SNE plots for read-count matrix imputed by FITs showed separability of minor populations cells, such as macrophages formed a group which was clearly visible as a separate group from other cell-types. We used CTS score to evaluate the separability of minor cell types after imputation. For minor cell types scImpute, DCA had noticeable CTS scores; however, they were ∼1.5 times lower than corresponding CTS values for FITs-based output (Supplementary Figure S4B).

Results for imputation of liver scATAC-seq profiles also revealed that when the read-count matrix was highly imbalanced, other methods (scImpute, MAGIC, KNNimpute and DCA) failed completely and could not help in segregating minor cells in t-SNE-based visualization (Figure 5A). On the other hand, FITs-based imputation caused the formation of separate groups for minor cell-types. There was a substantial difference between FITs and other four tested methods in terms of CTS (Figure 5B) and ARI score (Supplementary Figure S5 and Supplementary Table-1). Thus, the examples of bonemarrow and liver datasets, highlight the importance of sub-clustering using the tree-based approach in recovering signals of minor cells.

For FITs-based imputed version of the liver dataset, few cells with the label as 'unknown', co-localized with hepatocytes in t-SNE-based scatter plot (Figure 5A). We normalized the raw read-counts to highlight enhancers in 'Unknown' cells overlapping hepatocytes. We chose the top 10 000 potential enhancers based on the average of normalized read-count for 'unknown' cells co-localizing with hepatocytes and performed GREAT-based gene-ontology analysis (37). The top biological Process terms using GREAT-based analysis were related to functions of liver cells such as cholesterol and ketone metabolic processes. Thus it became quite evident that 'unknown' cells overlapping with hepatocytes were also typical liver cells. (Figure 5A and C). Using the same procedure on 'unknown cells' overlapping with endothelial-1 cells revealed top enriched gene ontology term related to endothelium development (Figure 5D). Thus, FITs-based imputation also enabled annotation of cells labeled as 'unknown'.

Further, we compared the performance of FITs with four other tools (chromVar (38), cisTopic (39), SCALE (19) and scOpen (20)) which were previously shown to be useful for visualization and clustering of scATAC-seq profiles. Using boneMarrow scATAC-seq profile, we realized that other four tools (chromVar, cisTopic, SCALE and scOpen) could not recover the signal of dendritic and T cells which had frequency <1.3% (Supplementary Figure S4C). With liver scATAC-seq profile, none of the four tested tools (chromVar, cisTopic, SCALE and scOpen) could display minority cell-types separately like FITs-based results (Figure 5E and Supplementary Figure S5C). With scATAC-seq profile of mouse cerebellum cells, the separability, according to cell-types in visualization and clustering purity was again better for FITs in comparison to chromVar, cisTopic, SCALE and scOpen (Supplementary Figure S6). Imputation by FITs also seemed to be useful for improving the performance of chormVar (Supplementary Figure S7). Further, we found that computation-time needed by FITs is similar to cisTopic and SCALE (Supplementary Table S2).

FITs improves the accuracy of cell-wise gene-set enrichment calculation and related analysis of atlas scale scATAC-seq profile

In addition to visualization and clustering, scATAC-seq profiles can also be used for many other purposes. Such as recently, Chawla *et al.* (40) developed a method called UniPath to transform scATAC-seq read-count to gene-set enrichment score for every single-cell, which can be used for inferring regulatory pattern in a cell. When gene-sets of cell-type markers are used, it is possible to annotate the cells using their scATAC-seq profile with UniPath (40). For estimating enrichment of gene-sets, UniPath first highlights cell-type specific peaks (possibly enhancers) by a division of read-count with pre-compiled global accessibility score. Then, it uses genes proximal to peaks with high cell-type specificity as a foreground to calculate gene-set scores. Hence, as shown above (Figure 3B), FITs improves the detection of cell-type specificity of peaks, it is highly likely that it would also improve the performance of UniPath.

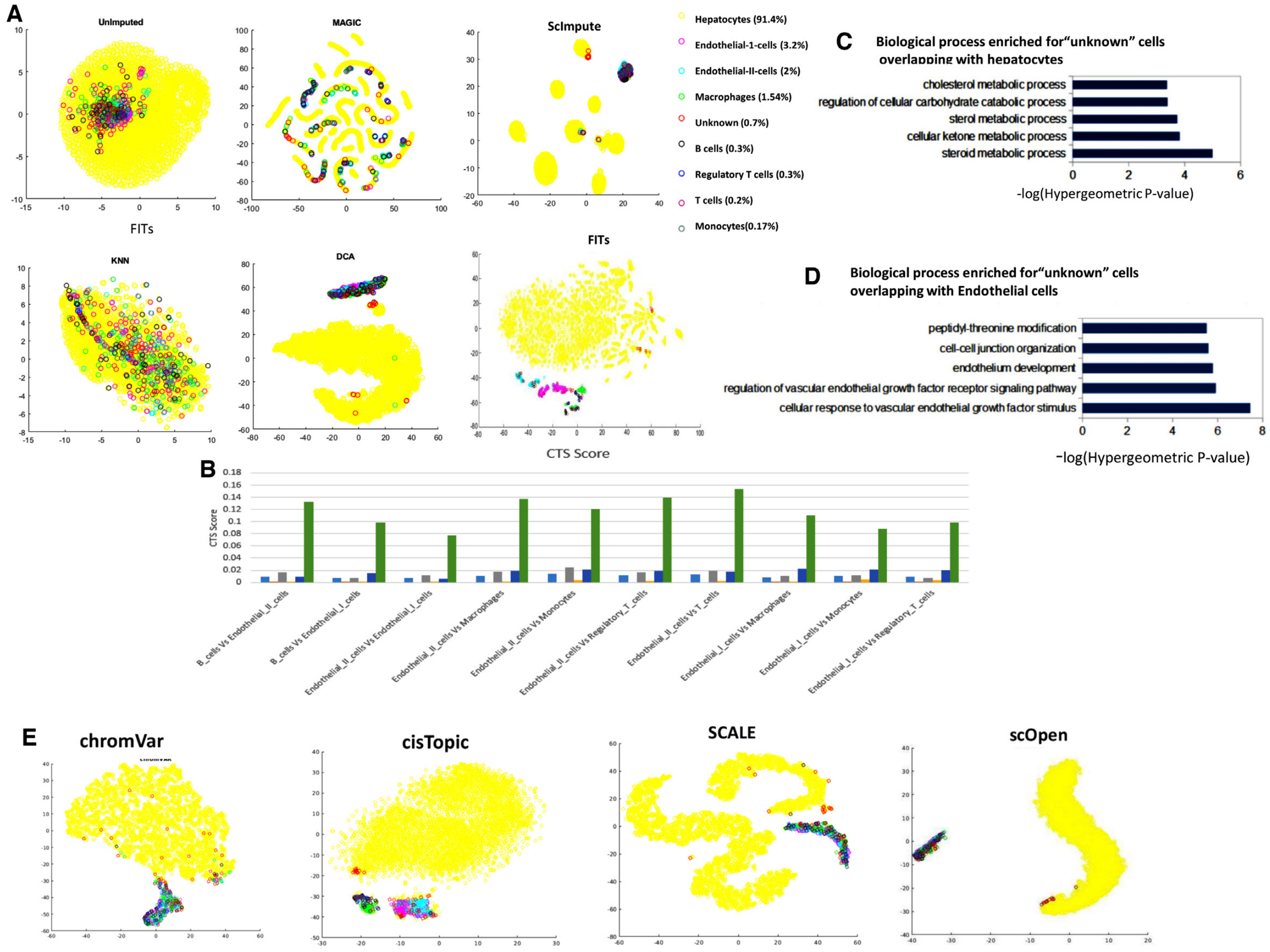

Figure 5. FITs recovers signal of minor cell-types in imbalanced single-cell ATAC-seq profile from *in vivo* sample. The dataset used here is the scATAC-seq read-count matrix for cells in adult mouse liver. (**A**) Scatter plot of t-SNE results for unimputed (raw) and read-counts matrix imputed by five methods. (**B**) CTS among different cell-types in read-count matrix imputed by five methods. (**C**) Top enriched gene-ontology terms (biological process) for predicted enhancers of 'unknown' cells co-localizing with hepatocytes in t-SNE-based plot for FITs output. For unbiased analysis, enhancers were predicted using unimputed read-counts of unknown cells co-localizing with hepatocytes. (**D**) Top biological process terms enriched for predicted enhancers in unknown cells co-localizing with endothelial-1 cells in results of FITs-based t-SNE plot. Again, only unimputed read-count were used to predict enhancers. (**E**) Scatter plot of t-SNE results for liver dataset made using four other tools (chromVar, cisTopic and SCALE, scOpen) designed for visualization of scATAC-seq profile.

Using gene-sets of known cell-type markers, We first evaluated how the output of UniPath can be improved with imputation by FITs. FITs-based imputation improved performance of UniPath substantially during estimation of gene-set enrichment score. As can be seen in Figure 6A for three sets of cells, the fraction of cells with correct cell-types in the top five terms is much higher with imputed read-count than with their unimputed version.

Chawla *et al.* (40) showed that the transformation of read-counts to pathway scores also provides an alternative approach of handling large scale scATAC-seq profile with consistency and horizontal scalability. However, Chawla *et al.* performed visualization of atlas scale data only for scRNA-seq profile but not for scATAC-seq profile. Hence we performed visualization of atlas scale scATAC-seq profile published by Cusanovich *et al.*, using the transformation of read-count to pathway scores for more than 68 000 cells (see Supplementary Methods). We applied t-SNE-based visualization using pathway scores calculated for unimputed and FITs-based imputed read-counts. It can be seen in Figure 6B, cells of the same type co-localized together in the t-SNE plot made using pathway enrichment score for imputed read-count matrix. Whereas, the t-SNE plot made using pathway enrichment scores from unimputed read-count showed high overlap among different cell-types. Hence our results show that FITs-based imputation can dramatically improve the analysis of atlas-scale scATAC-seq profiles using gene-set enrichment scores.

FITs improves detection of chromatin interaction from single-cell open chromatin profile

After evaluating FITs for improvement in calculating similarity among cells, we investigated whether imputation can help in estimating co-accessibility among sites. Recently, Pliner *et al.* (12) proposed that regions with high co-

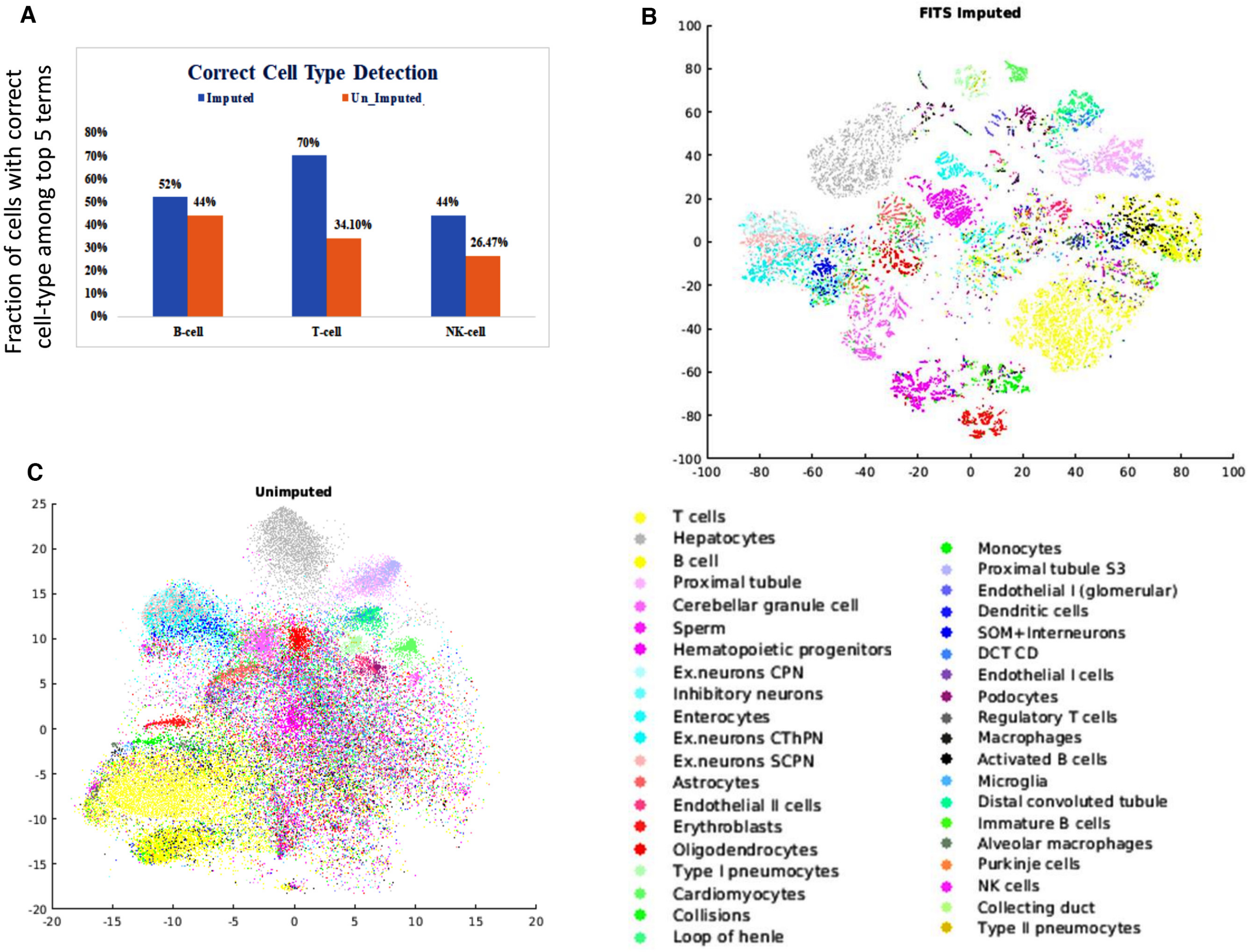

Figure 6. FITs improves analysis using gene-set enrichment score of each single-cell calculated using scATAC-seq profile. (**A**) The fraction of cells with correct cell-type terms appearing in top five enriched gene-sets, based on calculations using UniPath. FITs was used to impute the scATAC-seq profile of cells from Cusonovich *et al*. (6). (**B**) The t-SNE-based visualization using pathway scores calculated using read-counts imputed using FITs. (**C**) Visualization of t-SNE results for pathway scores calculated by UniPath using unimputed read-count.

accessibility in single-cell open chromatin profile are highly likely to be interacting. Chromatin interaction maps help in multiple processes such as identification of the target genes of noncoding genomic loci highlighted by GWAS (41) (Genome-wide association study) and understanding of gene regulation. Pliner *et al*. applied Graphical Lasso (26)-based approach to predict interaction among genomic sites. Even though the graphical Lasso method (26) is used to reduce the effect of noise and to calculate direct interactions, it's performance could be improved by providing less sparse data. We applied graphical Lasso-based approach to evaluate imputation-based improvement in the prediction of chromatin interaction using scATAC-seq. We used HiC-based chromatin interaction profile for K562 and GM12878, published by Rao *et al*. for evaluation (27). Using hic files we extracted high-confidence chromatin interactions at 25 Kbp resolution for both K562 and GM12878 cell lines. For both K562 and GM12878 cells, we merged the peaks lying within 25 Kbp in scATAC-seq read-counts matrix and applied Graphical Lasso to detect intrachromosomal chromatin interactions. For both cell types K562 and GM12878, FITs-based imputation indeed improved overlap among predicted and true high-confidence chromatin interaction by 10–30% for different chromosomes (Figure 7). One important issue to be noticed is that unlike Cicero, we did not focus on predicting interaction only within a certain distance range. Rather we also predicted intra-chromosome interaction between sites lying far apart. Thus, FITs prove to be useful for analysis of single-cell open-chromatin profile in multiple different ways, including chromatin interaction prediction.

DISCUSSION

The patterns of signal and sparsity in single-cell open chromatin are different in comparison to RNA-seq and DNA methylation profiles. Therefore, imputation of scATAC-seq profiles need attention, and it cannot be treated just like scRNA-seq dataset. Moreover, skipping imputation and doing binarization for scATAC-seq read-count cannot be fully justified. Even with a narrow peak of 200 bp, we can have read-count value as four read as each strand of two

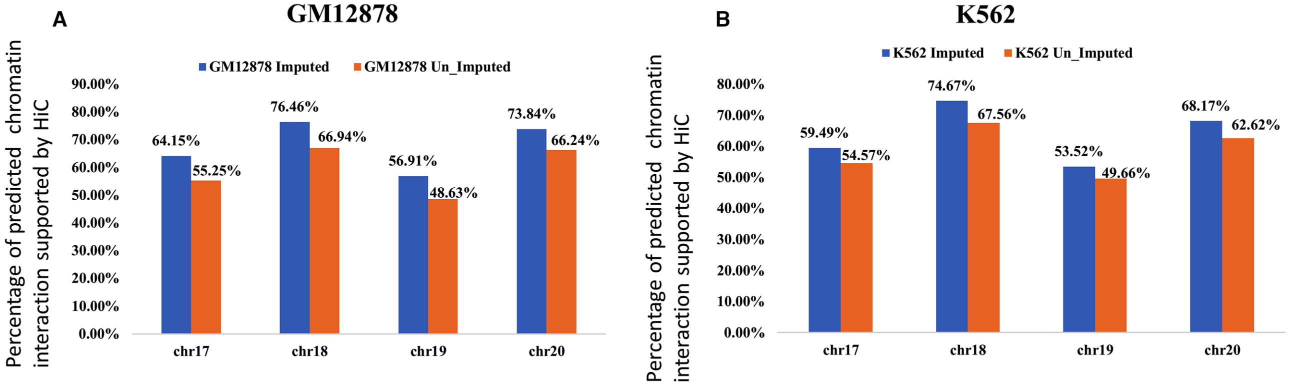

Figure 7. FITs-based imputation improves prediction of chromatin-interaction using scATAC-seq profile. The fraction of co-accessibility based predicted interaction overlapping with chromatin-interaction from HiC, is shown here. (**A**) for GM12878 cells (**B**) for K562 cells.

homologous chromosomes can contribute to DNA fragments pool. Moreover, highly active genomic sites tend to be bound by a large number of TFs causing wider region with chromatin-accessibility. Hence even if we use unique reads, we would have read-count value more than 1 for peaks of size >500 bp. Thus the read-count magnitude is indicative of activity-level of a genomic site which can be used to get insights about the effect of pathways and master regulators through noncoding regulatory sites with cell-type-specific activity. Hence cell-type specificity is a major concern while analyzing single-cell open-chromatin profiles. Moreover, there is a need to study the impact of imputation in different kinds of downstream analysis steps for single-cell open-chromatin profile. Therefore, first, we evaluated whether imputation of scATAC-seq can help in improving down-stream analysis desired for single-cell open-chromatin profile. Second, we developed a method for imputing which can handle a high level of noise, sparseness and cell-composition bias in single-cell open chromatin profiles. The strength of FITs lies in three features: randomized sub-clustering and imputation in multiple trees to avoid suboptimal solution, deciding drop-out after clustering and choosing imputed vectors based on correlation with the unimputed version to avoid wrong-imputation.

We have shown here that methods relying on parameters for single clustering step increase chances of artefacts due to errors in classification. Relying completely on few non-randomized classification steps also creates the risk of getting trapped in local minima and failure to detect true heterogeneity. We have shown here that FITs performs imputation in such a way that for scATAC-seq profile, there is less chance of detecting false heterogeneity in comparison to other imputation methods. Especially, when we have an unbalanced dataset, classification often fails to identify the minority population as a separate class, which creates artefacts during imputation by methods like scImpute and MAGIC. There have been multiple studies related to detecting rare cell-states using scRNA-seq profiles; however, with scATAC-seq such analysis is rarely done due to overwhelming noise and imbalance in datasets. Our analysis using FITs, hints that detecting rare cell-states using scATAC-seq read-counts is feasible, and it can provide a new direction in the analysis of clinical *in vivo* samples. For three scATAC-seq profiles of cells from *in vivo* samples, FITS showed better performance than four other methods (chromVar, cisTopic, SCALE and scOpen) designed for scATAC-seq profile. In addition to the recovery of minor-cell signals, we also showed the applicability of FITs for three analysis steps peculiar to scATAC-seq profiles which are; enhancer-detection, chromatin interaction prediction and calculation of gene-set enrichment score for single cells. Thus FITs can partner with many existing tools for improved inference and novel applications of scATAC-seq.

An advantage with FITs is that it can handle huge read-count matrices because of horizontal scalability. To run Phase-1 of FITs, one can break down the huge read-count matrix into many smaller matrices with randomly chosen cells. Two smaller matrices can have the same cell, but the union of all small matrices should represent all the cells in the original dataset. The Phase-1 of FITs can be run for multiple small matrices on different computers, before final matrix compilation by in Phase-2. Other imputation methods are rarely designed to handle huge read-count matrices. Hence FITs also resolves the problem of imputing large read-count matrices.

Multiomics studies using single-cell profile provide a global perspective of development and disease; however, very few groups have made such attempts (42). Thus, the major advantage of FITs is that it would encourage more researchers to explore single-cell open-chromatin profiles for multi-omics studies, due to reliability it adds during analysis. Other types of single-cell epigenome profiles such as histone modifications, MNAse-seq and DNAse-seq have also been used in few studies. The generality of FITs makes it suitable for other kinds of single-cell epigenome datasets also, therefore in future, FITs could be further adapted for other kinds of single-cell epigenome profiles.

The Python and Matlab version of FITs and https://reggenlab.github.io/FITs/ and imputed matrices used here for figures can be downloaded from http://reggen.iiitd.edu.in:1207/FITS/imputed_finaldata/.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

Authors contribution: V.K. and A.M. designed the study and reviewed the results and the manuscript. A.M., A.Mo., R.S. implemented the code in Matlab. S.M. implemented the code in python. V.K. arranged read-count for tested datasets and wrote the manuscript. R.S. and N.P. executed the programs in different datasets and created a few figures. N.P. also analyzed chromatin interaction profiles.

FUNDING

The computational resources and manpower were supported by intramural start-up grant from Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi) and Department of Biotechnology, India.
Conflict of interest statement. None declared.

References

1. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
2. Rivera,C.M. and Ren,B. (2013) Mapping human epigenomes. *Cell*, **155**, 39–55.
3. Kumar,V., Rayan,N.A., Muratani,M., Lim,S., Elanggovan,B., Xin,L., Lu,T., Makhija,H., Poschmann,J. and Lufkin,T. (2016) Comprehensive benchmarking reveals H2BK20 acetylation as a distinctive signature of cell-state-specific enhancers and promoters. *Genome Res.*, **26**, 612–623.
4. 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 Res.*, **23**, 2126–2135.
5. Rotem,A., Ram,O., Shoresh,N., Sperling,R.A., Goren,A., Weitz,D.A. and Bernstein,B.E. (2015) Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.*, **33**, 1165–1172.
6. 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. *et al.* (2018) A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*, **174**, 1309–1324.
7. Cao,J., Cusanovich,D.A., Ramani,V., Aghamirzaie,D., Pliner,H.A., Hill,A.J., Daza,R.M., McFaline-Figueroa,J.L., Packer,J.S., Christiansen,L. *et al.* (2018) Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, **361**, 1380–1385.
8. Jin,W., Tang,Q., Wan,M., Cui,K., Zhang,Y., Ren,G., Ni,B., Sklar,J., Przytycka,T.M., Childs,R. *et al.* (2015) Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature*, **528**, 142–146.
9. Lai,B., Gao,W., Cui,K., Xie,W., Tang,Q., Jin,W., Hu,G., Ni,B. and Zhao,K. (2018) Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature*, **562**, 281–285.
10. 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**, 486–490.
11. Jia,G., Preussner,J., Chen,X., Guenther,S., Yuan,X., Yekelchyk,M., Kuenne,C., Looso,M., Zhou,Y. and Teichmann,S. (2018) Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nat. Commun.*, **9**, 4877.
12. Pliner,H.A., Packer,J.S., McFaline-Figueroa,J.L., Cusanovich,D.A., Daza,R.M., Aghamirzaie,D., Srivatsan,S., Qiu,X., Jackson,D. and Minkina,A. (2018) Cicero predicts cis-regulatory DNA Interactions from single-cell chromatin accessibility data. *Mol. Cell*, **71**, 858–871.
13. Lareau,C.A., Duarte,F.M., Chew,J.G., Kartha,V.K., Burkett,Z.D., Kohlway,A.S., Pokholok,D., Aryee,M.J., Steemers,F.J. and Lebofsky,R. (2019) Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.*, **37**, 916–924.
14. van Dijk,D., Sharma,R., Nainys,J., Yim,K., Kathail,P., Carr,A.J., Burdziak,C., Moon,K.R., Chaffer,C.L., Pattabiraman,D. *et al.* (2018) Recovering gene interactions from single-cell data using data diffusion. *Cell*, **174**, 716–729.
15. Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Botstein,D. and Altman,R.B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.
16. Li,W.V. and Li,J.J. (2018) An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.*, **9**, 997.
17. Eraslan,G., Simon,L.M., Mircea,M., Mueller,N.S. and Theis,F.J. (2019) Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.*, **10**, 1–14.
18. Chen,H., Lareau,C., Andreani,T., Vinyard,M.E., Garcia,S.P., Clement,K., Andrade-Navarro,M.A., Buenrostro,J.D. and Pinello,L. (2019) Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.*, **20**, 241.
19. Xiong,L., Xu,K., Tian,K., Shao,Y., Tang,L., Gao,G., Zhang,M., Jiang,T. and Zhang,Q.C. (2019) SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.*, **10**, 4576.
20. Li,Z., Kuppe,C., Cheng,M., Menzel,S., Zenke,M., Kramann,R. and Costa,I.G. (2020) scOpen: chromatin-accessibility estimation of single-cell ATAC data. bioRxiv doi: https://doi.org/10.1101/865931, 05 November 2020, preprint: not peer reviewed.
21. Ji,Z., Zhou,W., Hou,W. and Ji,H. (2020) Single-cell ATAC-seq signal extraction and enhancement with SCATE. *Genome Biol.*, **21**, 161.
22. Candès,E.J. and Recht,B. (2009) Exact matrix completion via convex optimization. *Found. Comput. Math.*, **9**, 717.
23. Candes,E.J. and Plan,Y. (2010) Matrix completion with noise. *Proceedings of the IEEE*. Vol. **98**, pp. 925–936.
24. Cai,J.-F., Candès,E.J. and Shen,Z. (2010) A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.*, **20**, 1956–1982.
25. Li,C. and Zhou,H. (2017) svt: Singular value thresholding in MATLAB. *J. Stat. Softw.*, **81**, 1–13.
26. Friedman,J., Hastie,T. and Tibshirani,R. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.
27. Rao,S.S., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
28. Durand,N.C., Shamim,M.S., Machol,I., Rao,S.S., Huntley,M.H., Lander,E.S. and Aiden,E.L. (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.*, **3**, 95–98.
29. Greenwald,W.W., Li,H., Smith,E.N., Benaglio,P., Nariai,N. and Frazer,K.A. (2017) Pgltools: a genomic arithmetic tool suite for manipulation of Hi-C peak and other chromatin interaction data. *BMC Bioinformatics*, **18**, 207.
30. Van Der Maaten,L. (2014) Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.*, **15**, 3221–3245.
31. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
32. Buenrostro,J.D., Corces,M.R., Lareau,C.A., Wu,B., Schep,A.N., Aryee,M.J., Majeti,R., Chang,H.Y. and Greenleaf,W.J. (2018) Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, **173**, 1535–1548.
33. Teif,V.B., Mallm,J.P., Sharma,T., Mark Welch,D.B., Rippe,K., Eils,R., Langowski,J., Olins,A.L. and Olins,D.E. (2017) Nucleosome repositioning during differentiation of a human myeloid leukemia cell line. *Nucleus*, **8**, 188–204.
34. Chu,P.G. and Arber,D.A. (2001) CD79: a review. *Appl. Immunohistochem. Molecul. Morphol.*, **9**, 97–106.
35. Cusanovich,D.A., Reddington,J.P., Garfield,D.A., Daza,R.M., Aghamirzaie,D., Marco-Ferreres,R., Pliner,H.A., Christiansen,L., Qiu,X. and Steemers,F.J. (2018) The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*, **555**, 538–542.

36. Ng,A.Y., Jordan,M.I. and Weiss,Y. (2002) On spectral clustering: Analysis and an algorithm. *In Advances in neural information processing systems*. pp. 849–856.
37. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
38. Schep,A.N., Wu,B., Buenrostro,J.D. and Greenleaf,W.J. (2017) chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods*, **14**, 975–978.
39. Bravo Gonzalez-Blas,C., Minnoye,L., Papasokrati,D., Aibar,S., Hulselmans,G., Christiaens,V., Davie,K., Wouters,J. and Aerts,S. (2019) cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods*, **16**, 397–400.
40. Chawla,S., Samydurai,S., Kong,S.L., Wang,Z., Tam,W.L., Sengupta,D. and Kumar,V. (2019) UniPath: a uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles. bioRxiv doi: https://doi.org/10.1101/864389, 24 March 2020,preprint: not peer reviewed.
41. Zhang,F. and Lupski,J.R. (2015) Noncoding genetic variants in human disease. *Hum. Mol. Genet.*, **24**, R102–R110.
42. Lake,B.B., Chen,S., Sos,B.C., Fan,J., Kaeser,G.E., Yung,Y.C., Duong,T.E., Gao,D., Chun,J. and Kharchenko,P.V. (2018) Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.*, **36**, 70–80.