Gene Expression

Factorization-based Imputation of Expression in Single-cell Transciptomic Analysis (FIESTA) recovers Gene-Cell-State relationships

Elnaz Mirzaei Mehrabad [1,2], Aditya Bhaskara [2] and Benjamin T. Spike [1,2]*

[1]School of Computing, University of Utah, Salt Lake City, Utah, USA and
[2]Huntsman Cancer Institute, Department of Oncological Sciences, University of Utah School of Medicine, Salt Lake City, Utah, USA.

*To whom correspondence should be addressed.

Abstract

Motivation: Single cell RNA sequencing (scRNA-seq) is a powerful gene expression profiling technique that is presently revolutionizing the study of complex cellular systems in the biological sciences. Existing single-cell RNA-sequencing methods suffer from sub-optimal target recovery leading to inaccurate measurements including many false negatives. The resulting 'zero-inflated' data may confound data interpretation and visualization.
Results: Since cells have coherent phenotypes defined by conserved molecular circuitries (i.e. multiple gene products working together) and since similar cells utilize similar circuits, information about each each expression value or 'node' in a multi-cell, multi-gene scRNA-Seq data set is expected to also be predictable from other nodes in the data set. Based on this logic, several approaches have been proposed to impute missing values by extracting information from non-zero measurements in a data set. In this study, we applied non-negative matrix factorization approaches to a selection of published scRNASeq data sets to recommend new values where original measurements are likely to be inaccurate and where 'zero' measurements are predicted to be false negatives. The resulting imputed data model predicts novel cell type markers and expression patterns more closely matching gene expression values from orthogonal measurements and/or predicted literature than the values obtained from other previously published imputation approaches.
Contact: benjamin.spike@hci.utah.edu
Availability and implementation: FIESTA is written in R and is available at https://github.com/elnazmirzaei/FIESTA and https://github.com/TheSpikeLab/FIESTA.

1 Introduction

Single cell RNA sequencing (scRNA-seq) is a powerful laboratory technique aimed at quantifying the abundance of all the transcripts within individual cells. Although it is now a widely used approach for the identification of cell types and cell states based on characteristic gene expression patterns, scRNA-seq typically suffers from incomplete recovery of the cellular RNA pool within each cell (Marinov *et al.*, 2014; Linderman *et al.*, 2018; Huang *et al.*, 2018). Recently, several data imputation approaches have been proposed to address inaccuracy and 'zero-inflation' resulting from this transcript dropout effect (Huang *et al.*, 2018; van Dijk *et al.*, 2017; Li and Li, 2018; Linderman *et al.*, 2018; Arisdakessian *et al.*, 2019). Common to these computational approaches is the idea that missing values can be inferred and corrected by borrowing information from non-zero measurements obtained from similar cells and/or correlated genes. For example, the *scImpute* approach identifies similar cells by spectral clustering and then assigns a probability that a given zero value represents a dropout event and recommends a replacement value based on the bimodality and variance of expression distributions in closely clustering cells (Li and Li, 2018). Another approach, *MAGIC*, identifies similar cells using an adaptive Markov model in PCA space and subsequently imputes values for each gene using a diffusion model and pre-PCA values in 'neighboring' cells (van Dijk *et al.*, 2017). The *SAVER* approach assumes a negative binomial distribution of expressed genes and then uses the measured distributions from correlated genes and a penalized regression model to imply values where they are predicted to be missing (Huang *et al.*, 2018).

We hypothesized that a recommender system based on matrix factorization could provide a highly effective means to recover missing values in scRNA-Seq data since the approach has been widely used to make missing-value predictions from sparse data matrices in other fields (Cai *et al.*, 2010, Ling *et al.*, 2012, Dai *et al.*, 2020). Indeed, while we have been working to compile and vet the robust factorization-based computational pipeline reported here, a number of other approaches have been proposed that similarly rely on the principal of matrix factorization

to impute sparse and missing data. For example, *CMF-impute* uses a collaborative matrix factorization based on singular value decomposition (SVD) (Xu *et al.*, 2020). *ALRA* also employs SVD followed by a post hoc thresholding function (Linderman *et al.*, 2018). Even more similar to our own approach, the recently reported DeepImpute pipeline uses a deep neural network model to impute missing values (Arisdakessian *et al.*, 2019). Although perfect factorization of empirically derived matrices such as those containing gene expression data is NP-hard (Gillis and Glineur, 2011), machine learning techniques can be used to optimize the factorization, such that the product of factor matrix closely approximates the original matrix. This product matrix then serves as an idealization of the original matrix and can be used to recommend corrected values. NMF has been shown by several groups including our own to be particularly effective in delineating biologically relevant cell types and meaningful cell type associated gene expression profiles from cell expression data (Brunet *et al.*, 2004, Zhu *et al.*, 2017, Giraddi *et al.*, 2018). NMF thus represents an attractive approach to factorization-based imputation that is likely to draw from relevant biological substructures in the data.

The list of imputation methods described above is not exhaustive and several other approaches have also been reported (Ronen and Akalin, 2018; Xu *et al.*, 2020; Zand and Ruan, 2020; Chen and Zhou, 2018; Tang *et al.*, 2020; Lin *et al.*, 2017; and others) . These many attempts at accurately imputing missing data by borrowing information from the coherency of cell types and gene circuits attest to the widespread interest among sequencing users in the potential to computationally impute missing data that is biologically meaningful. Here, we present an unsupervised computational pipeline involving Factorization-based Imputation of Expression in Single-cell Transcriptomic Analysis (FIESTA), an NMF-based, machine-learning recommender system for imputation of missing values in scRNA-seq data. FIESTA is based on matrix reconstitution following either of two modified NMF approaches: sparse-NMF (sNMF) (Kim and Park, 2007) or Weighted NMF (WNMF) (Kim and Choi, 2009) and employs factorization rank, gene-weight, scaling and thresholding parameters derived from non-zero values in the original normalized matrix. We applied FIESTA to a selection of published data sets and compared its effectiveness to existing methods that are based on nearest-neighbor 'smoothing', as well as two recently reported matrix decomposition-based approaches: ALRA (Linderman *et al.*, 2018) and DeepImpute (Arisdakessian *et al.*, 2019). We find that FIESTA outperforms each of these techniques in recovering expression values known from orthogonal analysis of transcript levels or from the literature, is effective across a broader range of initial detection levels and facilitates the resolution of novel cell types and markers.

2 MATERIALS AND METHODS

2.1 Overview and input data

2.1.1 Pipeline overview

An overview of the FIESTA imputation package is shown in Figure 1. FIESTA first identifies critical parameters from the input matrix and then employs these parameters in the subsequent factorization, reconstitution and tempering (scaling/thresholding) steps to achieve an imputed data set with reasonably modeled expression values and greatly reduced false negative entries.

2.1.2 ScRNA-seq Datasets

The input for FIESTA is a scRNA-seq expression data set. scRNA-Seq data sets are essentially sparse/zero-inflated (m*n) matrices, where m is the number of genes and n is the number of cells in a given matrix, R. In this study, we used 3 different published datasets bearing orthogonal measurements and system knowledge from the literature that can serve as reasonable expectations of true expression values:

- a melanoma dropseq data set with paired orthogonal quantification of transcripts using in situ hybridization (Torre *et al.*, 2018),
- a recently published study of mouse lung adeno-carcinoma cells involving experimentally manipulated genetics and molecular therapeutic treatments, and bearing associated cell state changes (Zewdu *et al.*, 2021),
- and a portion (22184 genes * 3562 cells) of our previously published breast tissues scRNA-seq data set representing adult mouse mammary epithelial cells, where we have knowledge of tissue specific 'ground

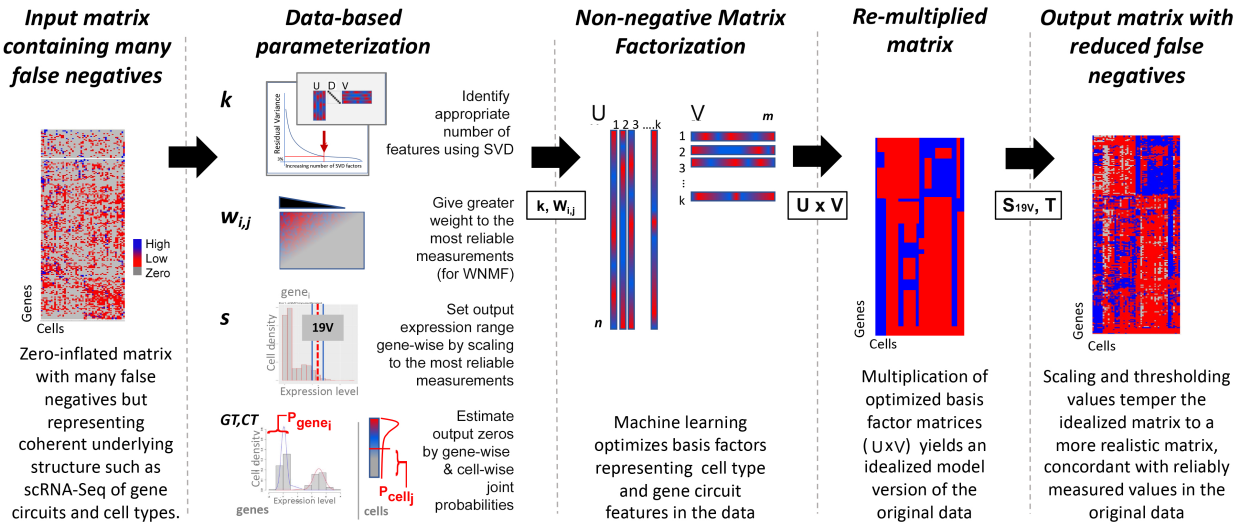


Fig. 1. Overview of FIESTA. FIESTA derives scaling, thresholding and factorization parameters from an input matrix, then factors the matrix with NMF, derives a factor product matrix and applies scaling and thresholding to create a tempered, imputed matrix.

truths' from our own studies and from the literature (Giraddi *et al.*, 2018).

2.1.3 Data pre-processing.

Raw data was independently normalized in each of the published studies we analyzes as described (Giraddi *et al.*, 2018, Torre *et al.*, 2018, Zewdu *et al.*, 2021). Logged values (i.e $Log_2(raw+1)$) were used as input for matrix factorization. Genes with no positive values in a data set are withheld from factorization and replaced in the output matrix as non-imputed.

2.2 Derivation of imputation parameters from input matrix

2.2.1 Finding an effective factorization rank (k).

We intend to use NMF to divide an expression data matrix into basis factors that we can subsequently multiply to generate an idealized matrix. This idealized product matrix can then be used to recommend corrected expression values. Ultimately, NMF factorization approaches have the advantage that basis factors are positive and therefore intuitively representative of cell types and gene expression circuits (Brunet *et al.*, 2004). We reasoned that the factorization rank (i.e. the number of features) should be tuned such that factors optimally mitigate noise but retain critical distinguishing features between similar cell types. Unfortunately, identification of an optimal k is NP-hard (Gheyas and Smith, 2010), and though iterative NMF-based optimization approaches can be effective they are usually slow. Singular Value Decomposition (SVD) is a much faster factorization approach and although it can produce uninterpretable basis vectors, studies have shown similar precision and recall for SVD and NMF at various ranks (k) (Peter *et al.*, 2009, Lawson and Hanson, 1995,Phillips *et al.*, 2009). Furthermore, since SVD also gives a 'variance capture' value from the diagonal matrix D, we used SVD to find a good approximation of number of features (k) in scRNAseq data and then used this k value for the rank of NMF.

Thus, to identify a suitable rank (k), we determine the number of sequential factors in an SVD diagonal matrix needed to describe at least 97% of the data variance. This value was arrived at empirically as (for instance) increasing k beyond this point showed diminishing returns in the mammary data set (Figure 2A,B)(Giraddi *et al.*, 2018). However, this parameter, designated variance capture (vc) remains selectable in the computational package.

We select k for each new input matrix R as follows: Consider D is the diagonal matrix from SVD of the input matrix, and n is the number of cells, then,

$$diffs[n] = \sum_{i=1}^{n} D[i] - D[i-1] \text{ ,then}$$
$$\text{find k such as } diff[k] \geq diff[n] \tag{1}$$

2.2.2 Calculate scaling coefficient (19V)

Similar to classically employed 'housekeeping genes', we previously employed the 19th ventile of expressed genes in scRNA-Seq data (i.e. values between the 90th and 95th percentile of expressed values) (Giraddi *et al.*, 2018) as a reliable scaling tether. This expression ventile encompasses multiple gene expression values near the upper end of the overall expression distribution, which are therefore less sensitive to dropout events than lower expression values, but which avoid extreme high-end outlier values such as erroneously high "jack-pot" genes (Marinov *et al.*, 2014). For use in gene-wise scaling in section 2.2.5 below, we calculated the mean gene-wise 19th ventiles for each gene as follows:

$$19V : \{\text{90th - 95th percentile mean } = mean((0.9*n),(0.95*n))\} \tag{2}$$

2.2.3 Predict the number of true negatives.

We used a normal mixture model fitting technique to estimate the expressed gene range and to predict the number of true zeros that would be present in an accurately imputed data matrix. In this regard, a normal-normal mixture model is fit to the expression values of each gene independently $\{gene_{i,1}, gene_{i,2}, ..., gene_{i,n}\}$ in R using the *mixtools* package (Benaglia *et al.*, 2009) and the distribution with the greater mean is determined to model positive expression. The distribution with the lower mean represents low or absent expression and is therefore used to calculate the likelihood of any node for that gene being a true negative. That is, the probability of zero in this distribution is used to determine the predicted number of true zeros across the entire data set for a specific gene.

Independently, for each cell, a negative binomial distribution is fit to the expression levels of all the genes expressed in that cell $\{cell_{j,1}, cell_{j,2}, ..., cell_{j,n}\}$ in R using the *fitdistrplus* package (Delignette-Muller *et al.*, 2015), and the probability of a true negative based on the cell-wise distribution of expression values across all genes is calculated.

Both the gene-wise and cell-wise zero-probabilities are multiplied by their respective vector lengths (i.e. m and n) to yield an expected number of zeroes in each cell and in each gene, independently. This dual thresholding pipeline is carried out as follow:

- Fit a mixture of two normal distribution to raw expression values for each gene

$$\text{Mixture Model} = \lambda_1 * norm(\mu_1, \sigma) + \lambda_2 * norm(\mu_2, \sigma) \tag{3}$$

  and calculate the probability of zero of the normal distribution with the smaller mean, (μ_{min}) in:

$$P_{gene_i}(0) = \lambda_{min} * pnorm(0, \mu_{min}, \sigma) \tag{4}$$

  Thus, we define $t_{gene_i}(0)$ representing the expected number of zeros based on gene expression values across the entire data set, by multiplying the $(P_{gene_i}(0))$ by the length of the gene row(n).

- A likelihood of zero is then also calculated on a cell-wise basis. We fit a negative binomial distribution to raw expression values found in each cell $\{gene_{1,i}, gene_{2,i}, ..., gene_{m,i}\}$, and find the probability of entries equaling zero for that cell. Multiplying the calculated probability of a zero by the length of the cell column (i.e number of genes, m) we calculate a cell-based- zero-expectation, $t_{cell_j}(0)$.

2.2.4 Imputation step.

Either of two NMF factorization implementations are used to factorize the gene-expression matrix using rank k. We implemented these approaches using the R library in NMF (Gaujoux and Seoighe, 2010).

- **Weighted non-negative Matrix Factorization (WNMF):** WNMF, also known as ls-nmf (least squares nonnegative matrix factorization) introduced by Guoli Wang et al., deploys uncertainty measurements of gene expressions into NMF updating steps (Wang *et al.*, 2006). WNMF gets a weight matrix as input to emphasize more reliable cells (m) in the factorization step.

  Given a non-negative matrix R_{m*n}, WNMF calculates 2 nonnegative factors U_{m*k} and V_{n*k}, which minimize:

$$cost(U,V) = 1/2 \sum_{i=1}^{m} \sum_{j=1}^{n} W_{ij}(R_{ij} - [UV^T]_{ij})^2 \tag{5}$$

- **Sparse Non-negative Matrix Factorization (sNMF):** Sparse Non-negative Matrix Factorization introduced by Hyunsoo Kim et. al (Kim and Park, 2007), uses alternating non-negativity-constrained least squares in the updating steps, and in each step sNMF keeps the sparsity of the factorized matrices. Given a non-negative matrix R_{m*n}, sNMF calculates 2 non-negative factors U_{m*k} and V_{n*k}, which minimize

$$cost(U,V) = 1/2\{||R_{ij} - [UV^T]_{ij}||_F^2 + \alpha||V||_F^2 + \sum_{i=1}^{m}\beta||U(i,:)||_1^2\} \quad (6)$$

This technique is known to work well on sparse datasets, which makes it especially suitable for scRNA-seq data.

In both approaches, nndsvd is used to generate the function seed (Boutsidis and Gallopoulos, 2008), and was effective in identifying a good initialization point as both functions converged after just 5 iterations.

After factorizing the gene-expression matrix with either of the above techniques, 2 non-negative factors are generated U_{m*k} and V_{n*k}, in order to generate the imputed matrix

$$ImputedR = U_{m*k} * (V_{n*k})^T \quad (7)$$

2.2.5 Scaling

We normalized the magnitude of imputed values in matrix ImputedR to the biological expectations on a gene-wise basis using the **19V** value determined from raw data above [2.2.2].

$$R_{SI}[gene_i,] = ImputedR[gene_i,] * S_{19V} \quad \text{where} \quad S_{19V} = (19V_{Raw}/19V_{ImputedR}) \quad (8)$$

2.2.6 Thresholding

To align the number of zeros in the scaled, imputed matrix R_{SI} with the cell-wise and gene-wise biological zero-expectations 9i.e. $t_{cell_j}(0)$ and $t_{gene_i}(0)$) described above [2.2.3], we calculate threshold vectors for all genes and cells (GT and CT, respectively), where:

$$GT = t_{gene_i}{}^{\text{th}} \text{ smallest values in } \{gene_{1,i}, ..., gene_{m,i}\} \text{ , for } \forall \text{ i}$$
$$CT = t_{cell_j}{}^{\text{th}} \text{ smallest values in } \{cell_{1,j}, ..., cell_{n,j}\} \text{ , for } \forall \text{ j} \quad (9)$$

An m*n thresholding matrix (T) is then created containing for each node the lesser of GT_i and CT_j. In the final step this threshold matrix T is applied to R_{SI} as follows:

$$R_{TSI_{ij}} = \{ \text{ if } R_{SI_{ij}} < T_{ij} \text{ ,then 0; else } R_{SI_{ij}}\} \quad (10)$$

That is, if the recommended value from the imputation step given in $R_{SI_{ij}}$ surpasses the minimum of these thresholds given in T_{ij}, the imputed value is retained. However, if the value given in $R_{SI_{ij}}$ is below the minimum of CT and GT thresholds given in T_{ij} that node will be zeroed out.

3 RESULTS

3.1 Imputation parameters derived from intrinsic properties of the data set.

Our approach to the recovery of missing values in a zero-inflated scRNASeq data begins with determination of a suitable number of factors to use in the matrix factorization step (Figure 1 and 2). These factors correspond to data features in the form of gene combinations whose expression distinguishes cell types and states (Brunet *et al.*, 2004). However, since the optimization of k in NMF is computationally intensive and slow using current techniques, we used SVD as a simple and rapid factorization approach to identify a suitable number of features (k) for subsequent use in NMF [see 2.2.1, above] (Phillips *et al.*, 2009). Thus, we calculated the percentage of total variance encompassed by sequentially increasing ranks of SVD for each filtered data set, independently. For example, at k=27 in the mammary epithelial data set (Giraddi *et al.*, 2018),

A

K	1	2	3	4	....	25	26	27	28	29	....
Coverage	0.7528	0.8446	0.9051	0.9144	...	0.9691	0.9695	**0.9705**	0.9709	0.971	...

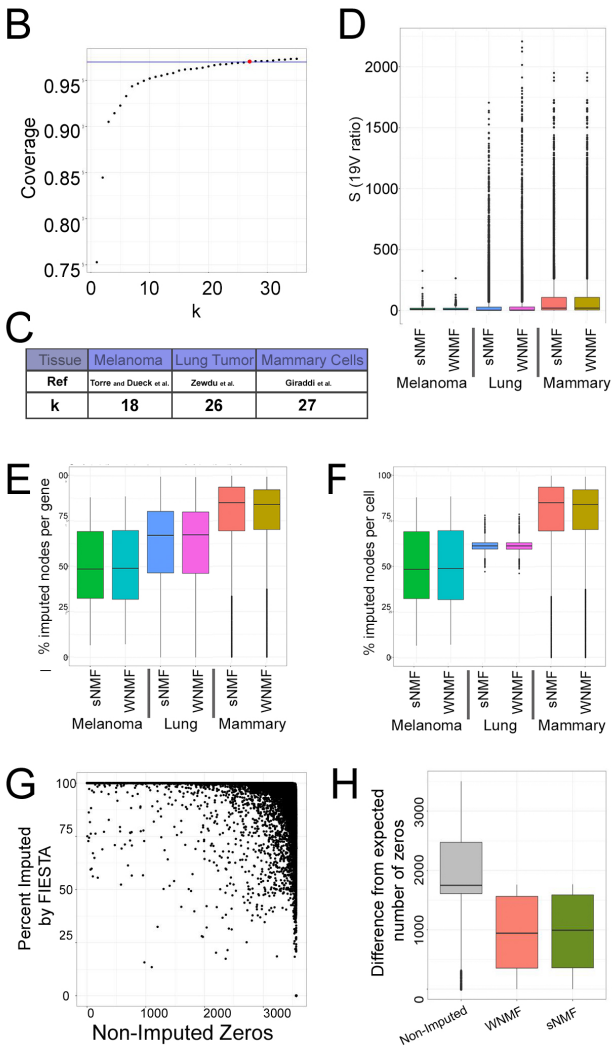


Fig. 2. Imputation, scaling and thresholding parameters. A) Identification of SVD diagonal matrix k surpassing 97% coverage of variance in Giraddi et al. data. B) Diminishing returns of increasing k in the Giraddi et al. data set. C) Unique threshold values for k in each data set. D) Gene-wise degree of scaling using S_{19V} (raw 19V-raw/19V-imputed) from sNMF and WNMF for each data set. E) Degree of gene-wise imputation varies by data set. F) Degree of cell-wise imputation varies by data set G) Percentage of zeros imputed vs number of zeros in the raw data in Giraddi et al. data. H) FIESTA imputation mitigates zero inflation in Giraddi et al. data.

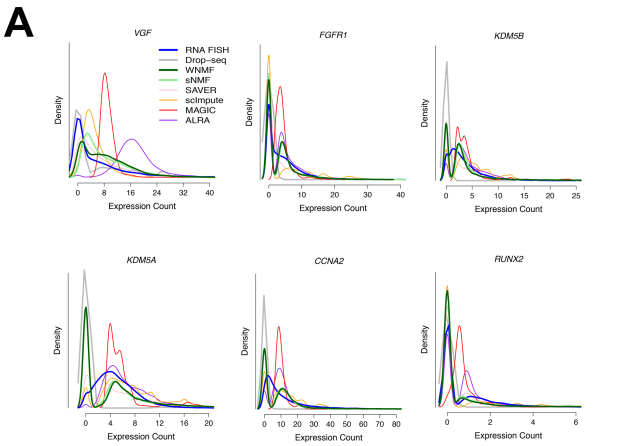
the SVD model captured >97% of the total variance and increasing k had diminishing returns (Figure 2A,B). We subsequently employed the 97% capture threshold for each data set we examined, although this parameter remains tunable by end users in the FIESTA package. The numbers of factors recommended for each data set analyzed in the present study based on >97% variance capture were 27 (Giraddi *et al.*, 2018), 26 (Zewdu *et al.*, 2021), and 18 (Torre *et al.*, 2018) (Figure 2C).

To derive scaling and thresholding coefficients for ensuring post-imputation matrices correspond to actual measured values where such values are reliable in the raw data, we first modeled the expression values for each gene using a mixture of two distributions [see Methods 2.2.3 and 2.2.6]. This mixed model fitting generates a predicted "expressed" and "low/not-expressed" distributions appropriate to the vast majority of genes. Next we determined the 19V of the upper, "expressed" distribution as a scaling factor for values resulting from the FIESTA computational pipeline [see Methods 2.2.2 and 2.2.5]. Thus, the 19V of the upper fit distribution divided by the 19V of imputed values on a gene-wise basis provides a scaling coefficient to bring imputed data to the scale of the most reliable measured values for each gene. While most scaling factors are modest, some usually those associated with low and rare expressed genes are larger (Figure 2D).

We also used the fit distributions to form a joint probability thresholding function to temper potential 'overimputation' of the data. We first calculated **1.** Probability of zero for each gene. To do this, a mixture of two normal distribution is fitted to each gene, and probability of zero in the normal distribution with lower mean is calculated as the gene-wise zero threshold (Figure 2) **2**. To calculate the probability of zero for each cell, a negative-binomial distribution is fitted to expressin values in each cell, then the probability of zero in the fitted distribution is considered as the cell-wise zero threshold. **3**.These unsupervised likelihood values were used to form joint probabilities that a given node equals zero rather than a recommended imputation value. **4**. For each node, if the imputed value of that node is bellow both gene-wise and cell-wise thresholds, the node will be zeroed out [see 2.2.3 and 2.2.6]. sNMF and WNMF performed comparably on each data set although WNMF was significantly faster (Figure 2,E,F and Supplemental Table 1). The overall amount of imputation was data set dependent and likely reflects sequencing depth with greater sequencing depth (i.e. Torre et al.>Zewdu et al.>Giraddi et al.) diminishing the amount of imputation carried out by FIESTA (Figure 2E,F). As a function of matrix factorization by parts (i.e. NMF) and gene specific thresholding, the proportion of zeros imputed was unique to each gene (Figure 2G). Even following the application of these scaling and thresholding steps which restore many zeros in the data, the imputation pipeline retained significant reduction in the number of zero values present in the expression matrix relative to the non-imputed matrix and this is consistent with the imputation of many predicted false negatives (Figure 2H).

3.2 FIESTA outperforms other imputation approaches across a broad range of expression values.

Single molecule fluorescence in situ hybridization (smFISH) provides an alternate approach to quantification of transcript abundance in individual cells. Torre and Dueck *et al.* measured the abundance of 15 distinct, variable transcripts in individual melanoma cells by smFISH that were also assessed by parallel scRNA-Seq from the same cell populations (Torre *et al.*, 2018). Not withstanding the challenges associated with enumerating smFISH data from a single confocal optical plain (Torre *et al.*, 2018), this orthogonal approach should provide a reasonable estimation of the true relative expression, and therefore a surrogate 'ground truth', as proposed in (Huang *et al.*, 2018). We therefore compared expression distributions measured by smFISH with the distribution of values recommended by FIESTA, in comparison to 4 alternative imputation techniques (Figure 3A,B).

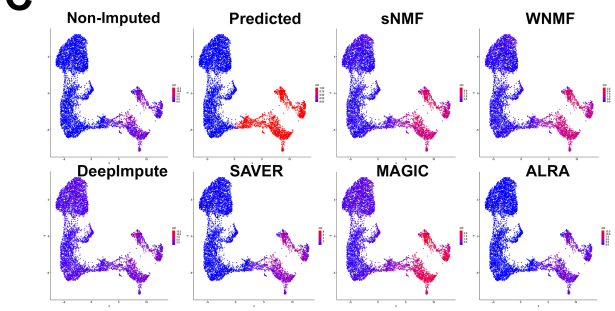


B

	FIESTA						
	sNMF	WNMF	ALRA	scImpute	MAGIC	Drop-Seq	SAVER
VCL	0.834	0.834	0.841	0.899	0.895	0.933	2.459
TXNRD1	0.865	0.869	0.892	0.94	0.912	0.923	1.343
MITF	0.844	0.848	0.876	0.868	0.873	0.886	0.93
CIS	0.931	0.933	0.943	0.966	1.495	0.989	2.337
BABAM1	0.79	0.791	0.792	0.884	0.846	0.94	1.541
FGFR1	0.577	0.571	0.589	0.744	0.757	0.867	2.445
VGF	0.4	0.354	0.68	0.618	0.731	0.647	0.588
LMNA	0.971	0.971	0.98	1.075	0.98	0.98	1.266
SOX10	0.97	0.971	0.968	1.069	0.977	0.983	2.777
KDM5B	0.588	0.592	0.467	0.807	0.709	0.817	0.89
KDM5A	0.681	0.684	0.633	0.85	0.784	0.837	0.909
CCNA2	0.697	0.704	0.649	0.822	0.775	0.899	1.082
JUN	0.48	0.465	0.585	0.699	0.39	0.843	3.451
RUNX2	0.43	0.436	0.382	0.349	0.689	0.657	1.826
FOSL1	0.682	0.707	0.733	0.463	0.918	0.106	0.699
Overall Rank	1	2	3	4	5	6	7

Hellinger Distance 0.000

RANK 1 2 3 4 5 6 7



D

Nkx2-1	RAW	WNMF	sNMF	MAGIC	SAVER	ALRA
Sensitivity	0.778	0.96	0.95	1	1	0.779
Specificity	0.879	0.978	0.979	NA	NA	0.879

Fig. 3. Performance of FIESTA in data recovery compared to existing methods. A) Select gene expression distributions obtained from Drop-Seq (i.e. non-imputed), smFISH, FIESTA, or alternative imputation approaches. B) Hellinger distance measures and ranking of imputation approaches. C) Predicted Nkx2.1 positive cells in Zewdu et al. based on gene deletion kinetics or various imputation approaches. D) Sensitivity and specificity measures for imputed Nkx2.1 in Zewdu et al.

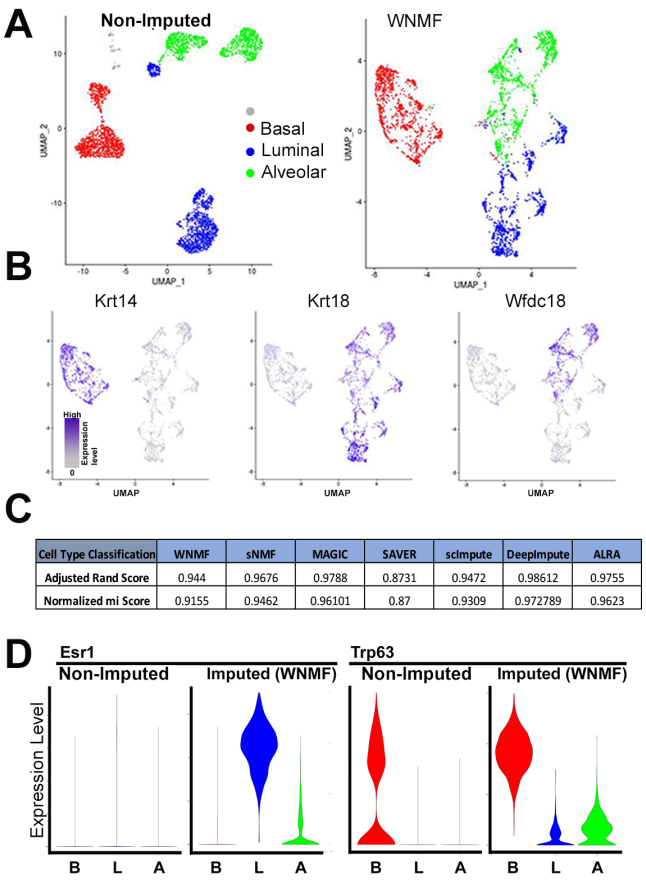
While each imputation approach increased the number of cells predicted to express the genes of interest, in most cases FIESTA imputation resulted in gene expression distributions better resembling those implied by smFISH analysis (Figure 3A,B). In fact, FIESTA (using either sNMF or wNMF approaches) yielded the top two closest approximation of smFISH data in over half of the genes tested and was within the top 3 scores (i.e. lowest 3 Hellinger distances) for all but 2 genes – Fosl1 which had negligible expression in this data set and Runx2 where FIESTA gave a close 3rd and 4th best approximation, behind scImpute and ALRA (Figure 3B and Supplemental Figure 1).

The lung tumor data set from Zewdu *et al.* also include a surrogate ground truth (Zewdu *et al.*, 2021). It involved experimentally induced genetic recombination to delete the Nkx2.1 gene, but gene deletion was only partly successful resulting in distinct cellular transcriptional profiles for Nkx2.1 deleted and non-deleted cells, implying expression and function of Nkx2.1 where deletion failed. We inferred that although Nkx2.1 was detected in only a subset of cells from the non-deleted clusters it was likely expressed but undetected in most cells sharing the phenotype of cells with detectable Nkx2.1 (Figure 3C)(Zewdu *et al.*, 2021). Using this inferred Nkx2.1 expression pattern, we tested the sensitivity and specificity of FIESTA and various alternative imputation approaches to the predicted expression pattern for the Nkx2.1 transcription factor in this study (Figure 3C,D). In this analysis SAVER and MAGIC exhibited apparent perfect sensitivity, but this derived from an overly liberal imputation of the data, as their specificity was incalculable due to no residual null values in the imputed data set (Figure 3D). In contrast, FIESTA scored highly in both sensitivity and specificity, while ALRA was overly conservative (Figure 3D).

3.3 Imputation affects cell-cell relationships

We next applied FIESTA to an adult subset of mouse mammary gland scRNASeq profiles that we published recently (Figures 4-7)(Giraddi *et al.*, 2018). For comparison, we imputed the same data set using other published approaches (Figure 4-6). In this mammary gland data set, we previously described three major cell types in detail based on transcriptional profiles and clustering. These cell types correspond to long established differentiated cells of the mammary gland and carry distinguishing marker gene expression including Krt14 for the basal cell type, Wfdc18 for the alveolar cell type and Krt18 (in the abscence of Wfdc18) for the remainder of luminal cells including the hormone sensing epithelial cells of the mammary gland (Figure 4A,B)(Giraddi *et al.*, 2018, Bach *et al.*, 2017).

Inspection of two-dimensional UMAP representations of imputed data compared to non-imputed data demonstrate that each imputation approach alters cell-cell relationships in unique ways (Figure 4A,B and Supplemental Figure 2). Although it is often assumed that measures of cluster separation or compactness (for instance in UMAP or tSNE plots) are useful measures of improved interpretability of the data following computational processing, there is usually no specific data supporting the conjecture that more discrete clusters are a more true representation of cellular relatedness, particularly since related cell types can lie along a continuum of phenotypic differences and even be somewhat interconvertible (Huang *et al.*, 2018, Regan and Smalley, 2020, Giraddi *et al.*, 2018). In this regard, we note that UMAPs derived from FIESTA imputation of mammary epithelial data yield a more continous luminal-alveolar relationship, but also yield substructures suggestive of cellular subtypes within the continuum of luminal/alveolar phenotypes, consistent with previous studies (Figure 4B)(Giraddi *et al.*, 2018, Bach *et al.*, 2017). Despite the changing graphical representations following imputation of the data, determination of adjusted Rand indices (ARI) demonstrate that all methods maintained the major cell type classifications in this data



Cell Type Classification	WNMF	sNMF	MAGIC	SAVER	scImpute	DeepImpute	ALRA
Adjusted Rand Score	0.944	0.9676	0.9788	0.8731	0.9472	0.98612	0.9755
Normalized mi Score	0.9155	0.9462	0.96101	0.87	0.9309	0.972789	0.9623

Fig. 4. Imputation alters apparent cell-cell relationships. A) UMAP of non-imputed mammary epithelial data and identification of basal, luminal and alveolar clusters based on expression of Krt14, Krt18 and Wfdc18. B) UMAPs of imputed data with expression patterns of cell type markers. C) Adjusted Rand scores for various imputation approaches. D) pre- and post-imputation values (FIESTA) for Esr1 and Trp63.

set (Figure 4A,B and Supplemental Figure 2). All approaches yielded ARI distances for imputed data >0.87 when using Seurat's graph-based clustering and marker genes (i.e. Krt14, Krt18, and Wfdc18) to call cell type clusters and then comparing imputed calls to non-imputed calls (Figure 4C)(Stuart *et al.*, 2019).

However, representation of known marker genes for some clusters were significantly improved following imputation. For instance, Krt18+/Wfdc18- cells show overly sparse expression of estrogen receptor though estrogen receptor is known to be broadly expressed by these cells and showed a generally active regulon score in previous studies (Bach *et al.*, 2017, Giraddi *et al.*, 2018, Zeps *et al.*, 1998) (Figure 4D). In contrast to non-imputed data, where Esr1 expression was rarely detected, FIESTA resulted in a model of Esr1 expression that was more broadly expressed but still compartmentalized in a manner consistent with the literature (Zeps *et al.*, 1998). The basal transcription factor Trp63 behaved similarly with the proportion of cells predicted to express Trp63 increasing after imputation but remaining largely basal restricted (Figure 4D).

3.4 Imputation affects gene-cell and gene-gene relationships

Although we also carried out orthogonal smFISH based validation of gene expression in Giraddi et al., the analysis was qualitative rather than

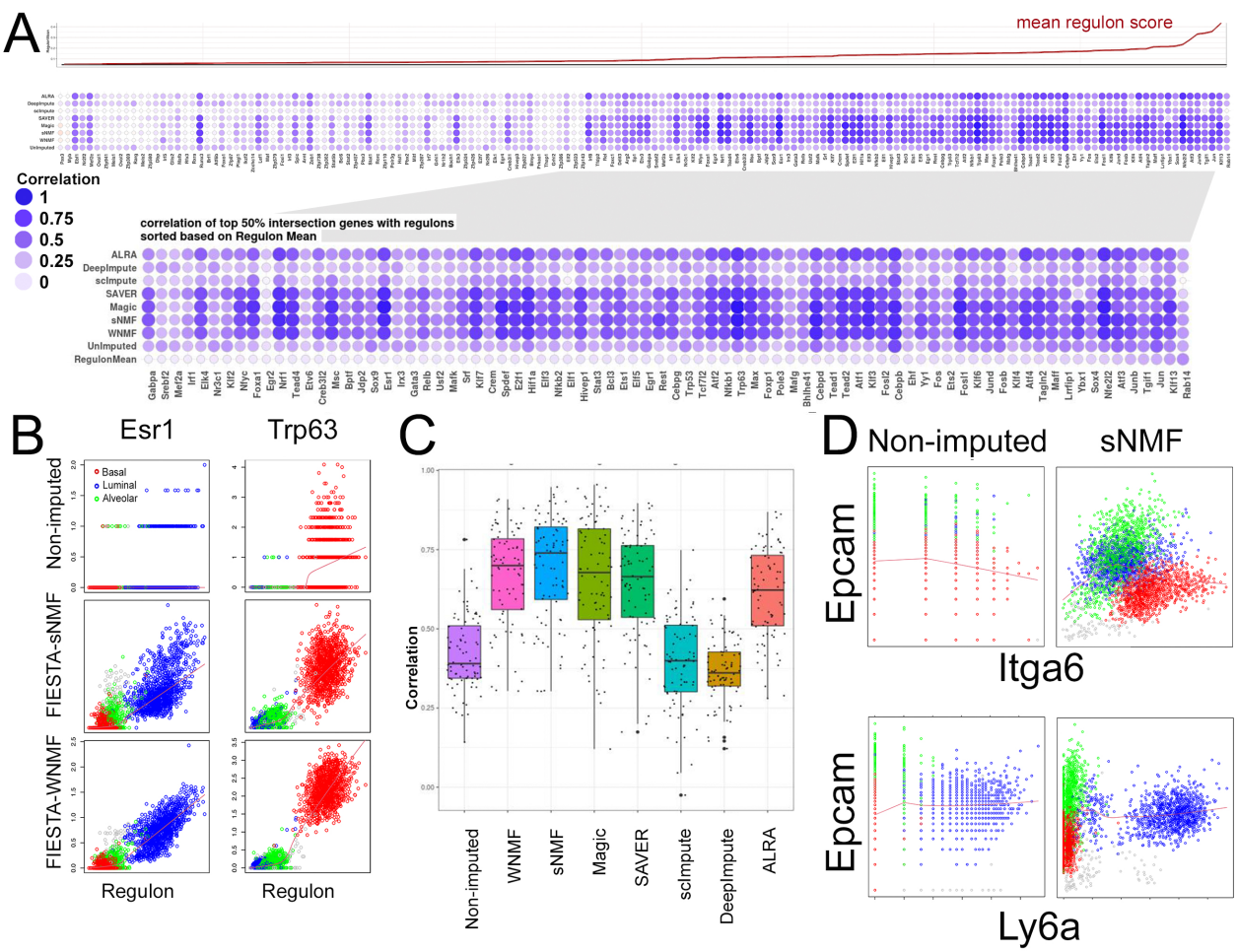


Fig. 5. Correlation of FIESTA imputation values with orthogonal gene expression predictions. A) Correlation of imputed transcription factor values with Regulon activity scores given by SCENIC in the Giraddi et al. B) The top half of regulon scores show high correlation with imputed values from several approaches including FIESTA. C) Correlation of Esr1 and Trp53 regulons with FIESTA predictions. D) FIESTA has the highest mean correlation with regulon predictions from SCENIC among imputation approaches. E) Correlations between FIESTA imputed values for Epcam and Ly6a or Itga6.

quantitative (Giraddi *et al.*, 2018). Thus, for this data set we leveraged a different orthogonal computational approach as a ground truth reference. Specifically, we compared values recommended by FIESTA (or other imputation approaches) with the results of regulon analysis that was carried out in Giraddi et al. using SCENIC (Aibar *et al.*, 2017, Giraddi *et al.*, 2018). SCENIC uses correlated genes across a data set as well as transcription factor binding site information to identify whether a transcription factor that is responsible for driving the expression of a set of genes is likely to be active. Where SCENIC predicts transcription factor activity it can be inferred that the transcription factor was present even if it was not detected in the raw scRNA-seq data. Thus, for a subset of genes (i.e. transcription factors) SCENIC operates as an imputation engine.

When we examined active regulons predicted by SCENIC and compared them to values recommended for the corresponding transcription factor from FIESTA or other approaches, we noted a strong concordance for many of the transcription factors in the top half of predicted activity scores (Figure 5A,B). For instance the luminal transcription factor Esr1 and the basal transcription factor Trp63 showed positive correlations between the SCENIC predicted activity level and the FIESTA predicted expression level, even though FIESTA did not rely upon target gene binding information or knowledge of its direct targets as surrogates for these predictions (Figure 5C). In the lower half of predicted regulon activity scores there was much less concordance (Figure 5A). This presumably reflects the expression but inactivity of some transcription factors. Compared to other methods, FIESTA demonstrated the highest mean correlation between regulon values and imputed transcription factor values among the top half of regulon scores i.e. those reflecting a transcription factor that is predicted to be present and active (Figure 5C).

Although transcript levels are not always expected to correlate perfectly with proteins levels given the existence of translational- and protein stability-control mechanisms, imputation of the adult mammary epithelial data with FIESTA yielded improvements of gene-gene correlation when judged against relationships of their gene products reported in the literature (Figure 5E and Supplemental Figure 3)(Visvader and Stingl, 2014; Shehata *et al.*, 2012; Asselin-Labat *et al.*, 2008; Spike *et al.*, 2012). For example, FIESTA imputation specifically recapitulates the general relationship between Epcam, Sca-1 (Ly6a) and CD49f (Itga6) proteins often used for sorting discrete mammary cell types (Figure 5D). Conservative imputation approaches left these genes and their cognate relationships highly digitized, while overly liberal approaches yielded relationships that (although possible) do not well match the relationships

described for their protein products in the literature (Supplemental Figure 3).

3.5 Imputation increases detection of population enriched marker genes and identifies cellular subtypes

We next asked whether running FIESTA on sparse scRNA-Seq data is likely to facilitate the identification of novel molecular markers for discrete cell types. Differential expression analysis using Wilcoxon Rank Sum in Seurat (Stuart *et al.*, 2019) identified 976, 724, and 297 differentially expressed genes uniquely overexpressed in basal, luminal and alveolar cellular subtypes, respectively (Supplemental Figure 4A). These genes represent many well known and often robustly expressed cell type identifiers in the mammary epithelium. However, following FIESTA, we identified an additional 1770, 1455, and 1103 candidate cell-type markers among the differentially expressed genes (Supplemental Figure 4A). Within this expanded gene list were genes with clearly cell-type restricted gene expression patterns but whose expression was detected only in very restricted number of cells in the raw data (Figure 6A,B and Supplemental Figure 4B). Thus, FIESTA helps identify sparsely represented cell type markers.

Finally, we asked whether apparent cellular subsets identified in the FIESTA imputed UMAPs were likely to reflect divergent cell types. We examined two presumptive alveolar and two presumptive luminal subsets for expression of additional transcripts related to known surface markers in the mammary gland. Interestingly, we identified distinguishable Egfr+Itga2+ and Kit+Itgb3+ positive subsets of Wfdc18+ cells, and also identified a Ly6a- subset of Esr1+ cells (Figure 6C,D). Although these cellular sub-states were not readily distinguishable in the non-imputed data, they may correspond to select cellular sub-states emerging from other recently published studies of cellular sub-types in the mammary gland (Regan and Smalley, 2020, Fu *et al.*, 2020, Pervolarakis *et al.*, 2020). Differential gene expression analysis between these subsets revealed the cells are likely distinguishable by broader transcriptomic differences representing distinct (although still quite similar) cell states (Figure 6D and Supplemental Table 2).

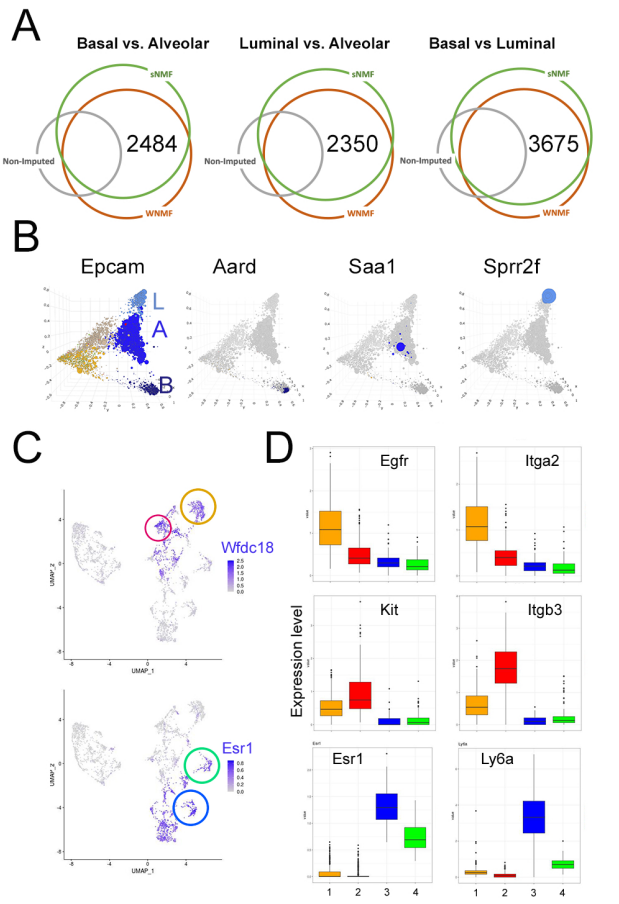


Fig. 6. Identification of novel cell type markers from imputed data. A) Differentially expressed genes identified in non-imputed and FIESTA imputed data. B) Diffusion map (as in Giraddi et al.) showing cell type specificity of imputed differentially expressed genes. C) Cellular subtypes implied by imputed data. D) Sub-population specific gene expression profiles of populations indentified in C.

4 Discussion

We set out to examine the potential of matrix factorization via NMF and subsequent matrix completion from factor multiplication to accurately impute missing values in scRNASeq data. This effort was based on the demonstrated effectiveness of NMF in identifying meaningful cell and gene relationships in gene expression data sets, and its proven utility in recommender systems dealing with other types of sparse data matrices. By applying unsupervised feature selection, scaling and thresholding parameters estimated from measured values in the raw data, we produced an unsupervised computational imputation pipeline for data processing, FIESTA, that outperforms several alternative approaches seeking to recover information in zero-inflated scRNA-Seq data. Further, by comparing results of this approach to expression patterns known from the literature and orthogonal analysis in three distinct published data sets, we found the approach produced reliable relative expression patterns for genes including those that had been undetected or poorly detected in the raw data. Although the large number of differentially expressed genes modeled by imputation of sparse scRNA-Seq data is likely to contain many false positives (Andrews and Hemberg, 2018), FIESTA permitted the identification of novel candidate cellular subtypes and markers for future study that were missed by the same analysis of non-imputed data. The approach will likely also have useful application to many of the emerging data sets and compendia that have been produced with the scRNA-Seq technology (e.g. Regev *et al.*, 2017, Consortium *et al.*, 2018).

While several of the previously proposed approaches also provide fairly reasonable gene expression estimates, these alternative methods may be overly conservative (e.g. Saver and ALRA), too liberal (e.g. MAGIC), or require potentially biased priors in the form of assumptions about 'important' genes or expected cell types (e.g. DeepImpute). In contrast FIESTA is unsupervised and gives expression values that realistically reflect both cellular heterogeneity and lineage/cell-type restriction. The initial version of FIESTA (version 1.0) as described in this manuscript is freely and immediately available as an R package [https://github.com/TheSpikeLab/FIESTA]. Subsequent studies are already underway to explore alternative NMF, scaling and thresholding algorithms that may further improve accuracy or reduce computational burdens, and be implemented as subsequent FIESTA version updates.

5 Summary

FIESTA, which is based on feature identification using wNMF or sNMF, imputes missing values in sc-RNASeq data with a more intuitive relationship to known biology than previously reported approaches. While, all approaches tested altered graphical representation of the data and thus affect interpretability of the data including cell-cell and cell-gene

relationships, the expression patterns following FIESTA worked across a broad range of input values and detection frequencies and matched values obtained from orthogonal approaches. Analysis of gene expression differences among populations emerging from FIESTA suggests resolution of biologically meaningful features and an enhanced ability to detect differentially expressed genes that define minority cell types.

Acknowledgements

The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged. We thank O.Balcioglu, D. Freeman and B. Gates for helpful discussion and careful reading of the manuscript.

Funding

This work has been supported by the Huntsman Cancer Foundation, and NIH Cancer Center Support Grant P30CA042014 funded cores at HCI, including the High-throughput Genomics and Bioinformatic core. The computational resources used were partially funded by the NIH Shared Instrumentation Grant 1S10OD021644-01A1.

References

Aibar, S., González-Blas, C. B., Moerman, T., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., *et al.* (2017). Scenic: single-cell regulatory network inference and clustering. *Nature methods*, **14**(11), 1083–1086.

Andrews, T. S. and Hemberg, M. (2018). False signals induced by single-cell imputation. *F1000Research*, **7**.

Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., and Garmire, L. X. (2019). Deepimpute: an accurate, fast, and scalable deep neural network method to impute single-cell rna-seq data. *Genome biology*, **20**(1), 1–14.

Asselin-Labat, M.-L., Vaillant, F., Shackleton, M., Bouras, T., Lindeman, G., and Visvader, J. (2008). Delineating the epithelial hierarchy in the mouse mammary gland. In *Cold Spring Harbor symposia on quantitative biology*, volume 73, pages 469–478. Cold Spring Harbor Laboratory Press.

Bach, K., Pensa, S., Grzelak, M., Hadfield, J., Adams, D. J., Marioni, J. C., and Khaled, W. T. (2017). Differentiation dynamics of mammary epithelial cells revealed by single-cell rna sequencing. *Nature communications*, **8**(1), 1–11.

Benaglia, T., Chauveau, D., Hunter, D., and Young, D. (2009). mixtools: An r package for analyzing finite mixture models. *Journal of Statistical Software*, **32**(6), 1–29.

Boutsidis, C. and Gallopoulos, E. (2008). Svd based initialization: A head start for nonnegative matrix factorization. *Pattern recognition*, **41**(4), 1350–1362.

Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, **101**(12), 4164–4169.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, **20**(4), 1956–1982.

Chen, M. and Zhou, X. (2018). Viper: variability-preserving imputation for accurate gene expression recovery in single-cell rna sequencing studies. *Genome biology*, **19**(1), 1–15.

Consortium, T. M. *et al.* (2018). Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, **562**(7727), 367–372.

Dai, X., Zhang, K., Xiong, J., Zhang, X., Tu, Z., and Zhang, N. (2020). Weighted non-negative matrix factorization for image recovery and representation. In *2020 10th International Conference on Information Science and Technology (ICIST)*, pages 288–293. IEEE.

Delignette-Muller, M. L., Dutang, C., *et al.* (2015). fitdistrplus: An r package for fitting distributions. *Journal of statistical software*, **64**(4), 1–34.

Fu, N. Y., Nolan, E., Lindeman, G. J., and Visvader, J. E. (2020). Stem cells and the differentiation hierarchy in mammary gland development. *Physiological reviews*, **100**(2), 489–523.

Gaujoux, R. and Seoighe, C. (2010). A flexible r package for nonnegative matrix factorization. *BMC bioinformatics*, **11**(1), 1–9.

Gheyas, I. A. and Smith, L. S. (2010). Feature subset selection in large dimensionality domains. *Pattern recognition*, **43**(1), 5–13.

Gillis, N. and Glineur, F. (2011). Low-rank matrix approximation with weights or missing data is np-hard. *SIAM Journal on Matrix Analysis and Applications*, **32**(4), 1149–1165.

Giraddi, R. R., Chung, C.-Y., Heinz, R. E., Balcioglu, O., Novotny, M., Trejo, C. L., Dravis, C., Hagos, B. M., Mehrabad, E. M., Rodewald, L. W., *et al.* (2018). Single-cell transcriptomes distinguish stem cell state changes and lineage specification programs in early mammary gland development. *Cell reports*, **24**(6), 1653–1666.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., and Zhang, N. R. (2018). Saver: gene expression recovery for single-cell rna sequencing. *Nature methods*, **15**(7), 539–542.

Kim, H. and Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, **23**(12), 1495–1502.

Kim, Y.-D. and Choi, S. (2009). Weighted nonnegative matrix factorization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1541–1544. IEEE.

Lawson, C. L. and Hanson, R. J. (1995). *Solving least squares problems*. SIAM.

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), 1–9.

Lin, P., Troup, M., and Ho, J. W. (2017). Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome biology*, **18**(1), 1–11.

Linderman, G. C., Zhao, J., and Kluger, Y. (2018). Zero-preserving imputation of scrna-seq data using low-rank approximation. *bioRxiv*, page 397588.

Ling, Q., Xu, Y., Yin, W., and Wen, Z. (2012). Decentralized low-rank matrix completion. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2925–2928. IEEE.

Marinov, G. K., Williams, B. A., McCue, K., Schroth, G. P., Gertz, J., Myers, R. M., and Wold, B. J. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and rna splicing. *Genome research*, **24**(3), 496–510.

Pervolarakis, N., Nguyen, Q. H., Williams, J., Gong, Y., Gutierrez, G., Sun, P., Jhutty, D., Zheng, G. X., Nemec, C. M., Dai, X., *et al.* (2020). Integrated single-cell transcriptomics and chromatin accessibility analysis reveals regulators of mammary epithelial cell identity. *Cell Reports*, **33**(3), 108273.

Peter, R., Shivapratap, G., Divya, G., and Soman, K. (2009). Evaluation of svd and nmf methods for latent semantic analysis. *International Journal of Recent Trends in Engineering*, **1**(3), 308.

Phillips, R. D., Watson, L. T., Wynne, R. H., and Blinn, C. E. (2009). Feature reduction using a singular value decomposition for the iterative guided spectral class rejection hybrid classifier. *ISPRS Journal of Photogrammetry and Remote Sensing*, **64**(1), 107–116.

Regan, J. L. and Smalley, M. J. (2020). Integrating single-cell rna-sequencing and functional assays to decipher mammary cell states and lineage hierarchies. *NPJ Breast Cancer*, **6**(1), 1–9.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., *et al.* (2017). Science forum: the human cell atlas. *Elife*, **6**, e27041.

Ronen, J. and Akalin, A. (2018). netsmooth: Network-smoothing based imputation for single cell rna-seq. *F1000Research*, **7**.

Shehata, M., Teschendorff, A., Sharp, G., Novcic, N., Russell, I. A., Avril, S., Prater, M., Eirew, P., Caldas, C., Watson, C. J., *et al.* (2012). Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. *Breast Cancer Research*, **14**(5), 1–19.

Spike, B. T., Engle, D. D., Lin, J. C., Cheung, S. K., La, J., and Wahl, G. M. (2012). A mammary stem cell population identified and characterized in late embryogenesis reveals similarities to human breast cancer. *Cell stem cell*, **10**(2), 183–197.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., III, W. M. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.

Tang, W., Bertaux, F., Thomas, P., Stefanelli, C., Saint, M., Marguerat, S., and Shahrezaei, V. (2020). baynorm: Bayesian gene expression recovery, imputation and normalization for single-cell rna-sequencing data. *Bioinformatics*, **36**(4), 1174–1181.

Torre, E., Dueck, H., Shaffer, S., Gospocic, J., Gupte, R., Bonasio, R., Kim, J., Murray, J., and Raj, A. (2018). Rare cell detection by single-cell rna sequencing as guided by single-molecule rna fish. *Cell systems*, **6**(2), 171–179.

van Dijk, D., Nainys, J., Sharma, R., Kaithail, P., Carr, A. J., Moon, K. R., Mazutis, L., Wolf, G., Krishnaswamy, S., and Pe'er, D. (2017). Magic: A diffusion-based imputation method reveals gene-gene interactions in single-cell rna-sequencing data. *BioRxiv*, page 111591.

Visvader, J. E. and Stingl, J. (2014). Mammary stem cells and the differentiation hierarchy: current status and perspectives. *Genes & development*, **28**(11), 1143–1158.

Wang, G., Kossenkov, A. V., and Ochs, M. F. (2006). Ls-nmf: a modified non-negative matrix factorization algorithm utilizing uncertainty estimates. *BMC bioinformatics*, **7**(1), 1–10.

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). Cmf-impute: an accurate imputation tool for single-cell rna-seq data. *Bioinformatics*, **36**(10), 3139–3147.

Zand, M. and Ruan, J. (2020). Network-based single-cell rna-seq data imputation enhances cell type identification. *Genes*, **11**(4), 377.

Zeps, N., Bentel, J. M., Papadimitriou, J. M., D'Antuono, M. F., and Dawkins, H. J. (1998). Estrogen receptor-negative epithelial cells in mouse mammary gland development and growth. *Differentiation*, **62**(5), 221–226.

Zewdu, R., Mehrabad, E. M., Ingram, K., Fang, P., Gillis, K. L., Camolotto, S. A., Orstad, G., Jones, A., Mendoza, M. C., Spike, B. T., *et al.* (2021). An nkx2-1/erk/wnt feedback loop modulates gastric identity and response to targeted therapy in lung adenocarcinoma. *Elife*, **10**, e66788.

Zhu, X., Ching, T., Pan, X., Weissman, S. M., and Garmire, L. (2017). Detecting heterogeneity in single-cell rna-seq data by non-negative matrix factorization. *PeerJ*, **5**, e2888.