Network-based imputation of dropouts in single-cell RNA sequencing data

Ana Carolina Leote[1,5,*], Xiaohui Wu[1,3,*], Andreas Beyer[1,2,4,5,#]

1. Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, 50923 Cologne, Germany
2. Cologne School for Computational Biology & Center for Data Science and Simulation, University of Cologne, 50923 Cologne, Germany
3. Department of Automation, Xiamen University, Xiamen 361005, China
4. Center for Molecular Medicine of the University of Cologne, 50931 Cologne, Germany
5. University of Cologne, Faculty of Medicine and University Hospital Cologne

[*] Equal contribution.

[#] Corresponding author:
Phone: +49-221-478 84429
Fax: +49-221-478 84026
E-Mail: andreas.beyer@uni-koeln.de

Abstract

Single-cell RNA sequencing (scRNA-seq) methods are typically unable to quantify the expression levels of all genes in a cell, creating a need for the computational prediction of missing values (‘dropout imputation’). Most existing dropout imputation methods are limited in the sense that they exclusively use the scRNA-seq dataset at hand and do not exploit external gene-gene relationship information.

Here, we show that a gene regulatory network learned from external, independent gene expression data improves dropout imputation. Using a variety of human scRNA-seq datasets we demonstrate that our network-based approach outperforms published state-of-the-art methods. The network-based approach performs particularly well for lowly expressed genes, including cell-type-specific transcriptional regulators. Additionally, we tested a baseline approach, where we imputed missing values using the sample-wide average expression of a gene. Unexpectedly, 52% to 77% of the genes were better predicted using this baseline approach, suggesting negligible cell-to-cell variation of expression levels for many genes. Our work shows that there is no single best imputation method; rather, the best method depends on gene-specific features, such as expression level and expression variation across cells. We thus implemented an R-package called ADImpute (available from http://cellnet.cecad.uni-koeln.de/adimpute) that determines the best imputation method for each gene in a dataset.

Introduction

Single-cell RNA sequencing (scRNA-seq) has become a routine method, revolutionizing our understanding of biological processes as diverse as tumor evolution, embryonic development, and ageing. However, the method still suffers from the problem that large numbers of genes remain undetected in single cells, although they actually are expressed (dropout events). Although dropouts are enriched among lowly expressed genes, relatively highly expressed genes can be affected as well. Of course, the dropout rate is also dependent on the sampling depth, i.e. the number of reads or transcript molecules (UMIs) quantified in a given cell. Imputing dropouts is necessary for fully resolving the molecular state of the given cell at the time of the measurement. In particular, genes with regulatory functions - e.g. transcription factors, kinases, regulatory ncRNAs - are typically lowly expressed and hence particularly prone to be missed in scRNA-seq experiments. This poses problems for the interpretation of the experiments if one aims at understanding the regulatory processes responsible for the transcriptional makeup of the given cell.

A range of computational methods have been developed to impute dropouts using the expression levels of detected genes. The underlying (explicit or implicit) assumption is often that detected and undetected genes are subject to the same regulatory processes, and hence detected genes can serve as a kind of 'fingerprint' of the state at which the cell was at the time of lysis. Most successful methods are based on some type of grouping (clustering) of cells based on the similarity of their expression patterns. Missing values are then imputed as a (weighted) average across those similar cells where the respective gene was detected[1–4]. These methods rest on two important assumptions: (1) the global expression pattern of a cell (i.e. across the subset of detected genes) is predictive for all genes; (2) the (weighted) average of co-clustering (i.e. similar) cells is a good estimator of the missing value. The first assumption is violated if the expression of a dropout gene is driven by only a small subset of genes and hence the global expression pattern does not accurately reflect the state of the relevant sub-network. Any global similarity measure of the whole transcriptome will be dominated by the majority of genes. The second assumption is violated if the data is scarce, i.e. when either only few similar cells were measured or if the particular gene was detected in only a small subset of cells. In that case the average is computed across a relatively small number of observations and hence unstable.

The recently published SAVER[5] method employs a different strategy that can overcome some of these limitations. SAVER attempts to predict the posterior probability distribution for the ‘true’ expression of each gene in each cell. This distribution is derived from the data by learning gene-gene relationships from the dataset, which are subsequently used to predict the gene- and cell-specific probability distribution of the expected expression value. In other words, SAVER does not use the whole transcriptome of a cell to predict the expression level of a given gene; instead, it uses a specific subset of genes that are expected to be predictive for the particular gene at hand.

Here, we compare published approaches that are representative for current state-of-the-art methods to two fundamentally different approaches. The first is a very simple baseline method that we use as a reference approach: we estimate missing values as the average of the expression level of the given gene across all cells in the dataset where the respective gene was

detected. Initially intended to serve just as a reference for minimal expected performance, this sample-wide averaging approach turned out to perform surprisingly well and in many instances even better than state-of-the-art methods. The simple explanation is that estimating the average using all cells is a much more robust estimator of the true mean than using only a small set of similar cells, especially when the gene was detected in only few cells and/or if the gene's expression does not vary much across cells.

The second new approach avoids using a global similarity measure comparing entire transcriptomes. Instead, similar to SAVER it rests on the notion that genes are part of regulatory networks and only a small set of correlated or functionally associated genes should be used to predict the state of undetected genes. However, unlike SAVER, we propose to use transcriptional regulatory networks trained on independent (bulk seq) data to rigorously quantify the transcriptional relationships between genes. Missing values are then imputed using the expression states of linked genes in the network and exploiting the known quantitative relationships between genes. This approach allows imputing missing states of genes even in cases where the respective gene was not detected in any cell or in only extremely few cells. This second new approach rests on the assumption that the network describes the true regulatory relationships in the cells at hand with sufficient accuracy. Here, we show that this is indeed the case and that combining the two new approaches with published state-of-the-art methods drastically improves the imputation of scRNA-seq dropouts. Importantly, the performance of an imputation method is dependent on the 'character' of a gene (e.g. its expression level or the variability of expression between cells). Hence, we implemented an R-package (Adaptive Dropout Imputer, or ADImpute) that determines the best imputation method for each gene through a cross-validation approach.

Results

Imputing dropouts using a transcriptional regulatory network

In order to understand whether the inclusion of external gene regulatory information allows for more accurate scRNA-seq dropout imputation, we derived a regulatory network from bulk gene expression data in 1,376 cancer cell lines with known karyotypes. For this purpose, we modelled the change (compared to average across all samples) of each gene as a function of its own copy number state and changes in predictive genes:

$$y_i = \alpha_i \cdot c_i + \sum_{j \neq i} \alpha_{ij} \cdot y_j + \varepsilon_i \quad , \qquad (1)$$

where y_i is the expression deviation (log fold change) of gene i from the global average, c_i is the known (measured) copy number state of gene i, α the vector of regression coefficients, y_j the observed change in expression of gene j and ε_i the i.i.d. error of the model. To estimate a set of predictive genes *j*, we made use of LASSO regression[6], which penalizes the L1 norm of the regression coefficients to determine a sparse solution. LASSO was combined with stability selection[7] to further restrict the set of predictive genes to stable variables and to control the false discovery rate (Methods). Using the training data, models were fit for 24,641 genes, including 3,696 non-coding genes. The copy number state was only used during the training of the model,

since copy number alterations are frequent in cancer and can influence the expression of affected genes. If copy number states are known, they can of course also be used during the dropout imputation phase. Using cell line data for the model training has the advantage that the within-sample heterogeneity is much smaller than in tissue-based samples[8]. However, in order to evaluate the general applicability of the model across a wide range of conditions, we validated its predictive power on a diverse set of tissue-based bulk-seq expression datasets from the The Cancer Genome Atlas (4,548 samples from 13 different cohorts; see Methods and Supplementary Figure S1).

Such a model allows us to estimate the expression of a gene that is not quantified in a given cell based on the expression of its predictors in the same cell. Here, the difficulty lies in the fact that imputed dropout genes might themselves be predictors for other dropout genes, i.e. the imputed expression of one gene might depend on the imputed expression of another gene. In order to derive the imputation scheme based on the model from equation (1), we revert to an algebraic expression of the problem,

$$Y = AY \,, \qquad (2)$$

where A is the adjacency matrix of the transcriptional network, with its entries α_{ij} being fitted using the regression approach described above, and Y is the vector of gene expression deviations from the mean across all cells in a given cell. In the current implementation we assume no copy number changes and hence, we exclude the c_i term from equation (1). Like in equation (1), we omit the intercept since we are predicting the deviation from the mean. Further note that we drop the error term ε from equation (1), because this is now a prediction task (and not a regression). Here, we exclusively aim to predict dropout values, and (unlike SAVER) our goal is not to improve measured gene expression values. Hence, measured values remain unchanged. It is therefore convenient to further split Y into two sub-vectors Y^m and Y^n, representing the measured and non-measured expression levels, respectively. Likewise, A is reduced to the rows corresponding to non-measured expression levels and split into A^m (dimensionality $|n| \times |m|$) and A^n (dimensionality $|n| \times |n|$), accounting for the contributions of measured and non-measured genes, respectively. The imputation problem is then reduced to:

$$Y^n = A^n Y^n + A^m Y^m \qquad (3)$$

As Y^m is known (measured) and will not be updated by our imputation procedure, the last term can be condensed in a fixed contribution, $F = A^m Y^m$, accounting for measured predictors:

$$Y^n = A^n Y^n + F \qquad (4)$$

The solution Y^n for this problem is given by:

$$Y^n = (I - A^n)^{-1} F \qquad (5)$$

The matrix $(I - A^n)$ may not be invertible, or if it is invertible, the inverse may be unstable. Therefore, we computed the pseudoinverse $(I - A^n)^+$ using the Moore-Penrose inversion. Computing this pseudoinverse for every cell is a computationally expensive operation. Thus, we implemented an additional algorithm finding a solution in an iterative manner (Methods). Although this iterative second approach is not guaranteed to converge, it did work well in

practice (see Supplementary Figure S2, Methods). While our R-package implements both approaches, subsequent results are based on the iterative procedure.

Gene regulatory network information improves scRNA-seq dropout imputation

To assess the performance of our network-based imputation method and compare it to that of previously published methods, we considered three different single-cell RNA sequencing datasets (Supplementary Table S1). The first dataset focuses on human embryonic stem cell (hESC) differentiation and comprises 1,018 cells − hESCs, lineage progenitors derived from hESCs and human foreskin fibroblasts[9]. Cell type labels are known for this dataset. A second dataset comprises 4,347 cells from 6 human oligodendrogliomas[10]. Finally, data from healthy pancreata of 2 human donors[11] were also used for test purposes. It was important to include a range of different healthy cell types in the evaluation, because the transcriptional regulatory network was trained on cancer cell line data. Thus, by including data from non-cancerous tissues, we could evaluate possible restrictions induced by the model training data.

In order to quantify the performance of both proposed and previously published imputation methods, we randomly set a fraction of the quantified values in the test data to zero (Methods) and stored the original values for later comparison with the imputed values. This masking of measured values largely maintains gene- and cell-specific biases, as the masked values are added to the zero entries originally present in the data. Imputation was then performed on the masked dataset using our network-based approach, scImpute[2] and SAVER[5]. Those two methods were chosen since they were shown to be among the top-performing state-of-the-art dropout imputation methods[12]. As a baseline method, the masked and actual dropout values were assigned the average $\log_2$TPM across all cells where the corresponding gene was quantified.

As expected, imputation error increased with increasing missing information (NAs) per gene in the data (Fig. 1). While this is true for all methods, scImpute (Fig. 1, green line) was particularly sensitive to missing information about genes across cells, a behaviour also described elsewhere[12]. Our network-based approach outperformed both scImpute and SAVER with increasing zero entries per gene (Fig. 1). As the network-based approach uses information regarding other genes contained in the same cell, we hypothesized its accuracy might be more affected by increasingly sparse information per cell when compared to other methods. However, although the average imputation error slightly increased with the number of missing genes per cell, this was true for all methods and the relative performance differences were largely invariant to the number of missing genes per cell (Fig. S3). Further, the network-based method performed well over a range of different cell types. Thus, the diversity of cell lines used in the training data seemed to capture a large fraction of all possible regulatory relationships in the human transcriptome.

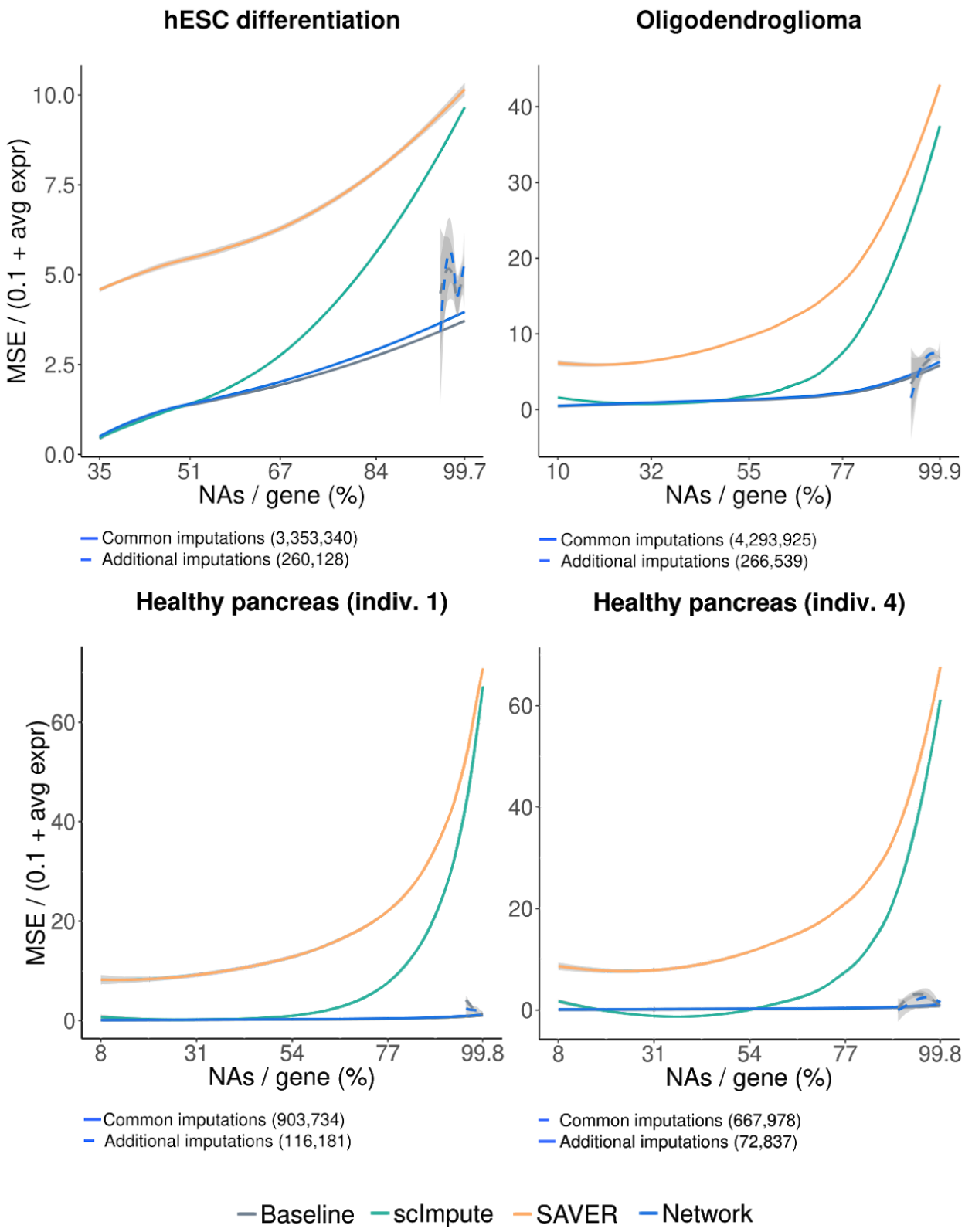

Figure 1: Imputation error of Baseline (slate grey), scImpute (green), SAVER (yellow) and Network (blue). Loess trendline of weighted Mean Squared Error (MSE) of imputation with percentage of missing values per gene, for all 4 test datasets, restricted to values that could be imputed by all four methods (full lines) and additional values imputed exclusively by both the baseline (dashed grey) and network (dashed blue) methods.

The performance of the baseline method (Fig. 1), which does not account for any expression variation between cells, was surprising to us. Unexpectedly, the expression of more than half of the genes was best predicted using the baseline method (Table 1). While SAVER showed a poor performance in comparison to all other methods (also described in [12]), it should be noted that this method aims to estimate the true value for all genes, not only for the dropout genes. Hence, its goal is slightly different from that of the other methods in this comparison.

Both the network-based and the baseline methods were able to perform imputations for genes for which SAVER and scImpute could not make any predictions. Notably, the quality of these extra imputations was comparable to that of imputations for genes commonly imputed by all methods (Fig. 1, dashed lines). The additional imputations were particularly enriched in rarely quantified genes, including those not quantified in any cell. In those cases only the network-based approach − which does not rely on the quantification of a gene in other cells − was able to perform non-zero imputations. Further, the performance of the network method for extremely rare genes was superior to that of the baseline (Fig. S4).

As an alternative way of assessing the performance of the imputation methods we quantified the correlation between the original values before masking and the results of each imputation procedure (Supplementary Table S2). As opposed to computing the residuals between imputed and measured values, the regression is independent of mean shifts in the predictions. We observed the same relative performance of the methods as with the imputation error analysis.

Taken together, these results indicate that both the baseline and the network-based approach often lead to more accurate and numerous imputations than state-of-the-art imputation methods.

Gene features determine the best performing imputation method

To characterize the genes best imputed by each of the methods, we determined, for each gene in each test dataset, the method resulting in the lowest weighted Mean Squared Error of imputation (Table 1).

Table 1: Percentage of genes best imputed by each method (lowest weighted MSE) in the four test datasets restricted to values that could be imputed by all four methods.

	Baseline	**scImpute**	**SAVER**	**Network**
hESC differentiation	50.2%	27.6%	0.0%	22.2%
Oligodendroglioma	79.5%	2.6%	0.0%	17.9%
Healthy pancreas (indiv. 1)	61.7%	0.0%	0.0%	38.3%
Healthy pancreas (indiv. 4)	66.8%	0.0%	0.0%	33.2%

The genes best imputed by each method were then compared against a background including all genes imputed by all 4 methods (Fig. 2 and Fig. S5). As expected, rarely detected genes were best imputed by the network and baseline methods. While moderately expressed, more

frequently detected genes were preferably imputed by scImpute, the network-based approach performed better for highly expressed and variable genes.

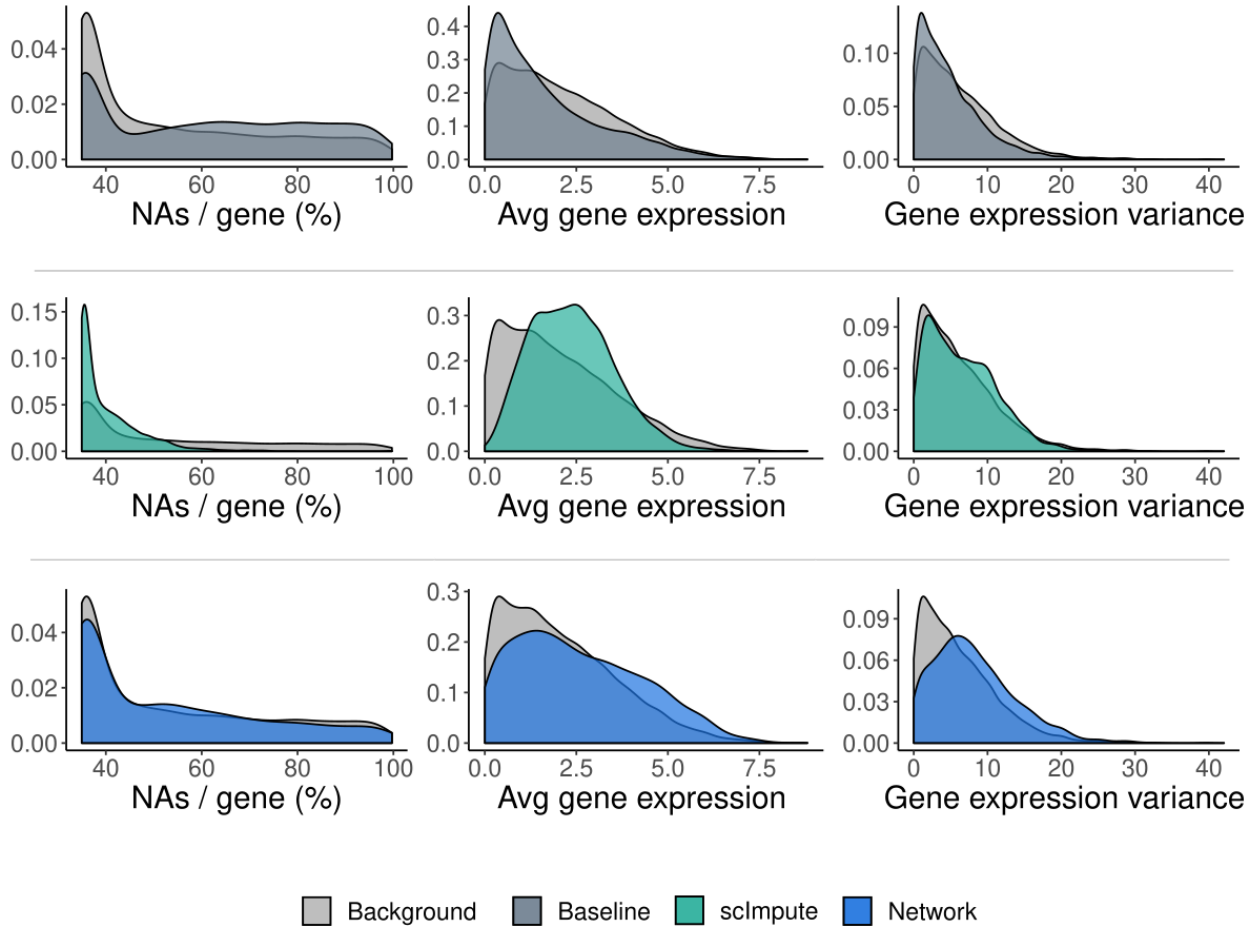

Figure 2: Characterization of the genes best predicted by the methods Baseline (slate grey), scImpute (green) or Network (blue) in the hESC differentiation dataset. Distribution of missing values per gene, average expression levels and variance of the genes best predicted by Baseline, scImpute and Network methods, compared against all tested genes (background). Due to the very low numbers of genes best imputed by SAVER, the corresponding distributions are not shown. Average gene expression is shown as $\log_2$-transformed TPM.

We next wondered if genes best predicted by a particular method were enriched for specific functions. Thus, we performed GO (Gene Ontology) enrichment analysis on the sets of genes best predicted by a given method, which did indeed yield significant enrichments, especially for the genes best imputed by the network-based approach (Figure S6). In particular, we noticed a trend for genes associated to mRNA- and protein-binding molecular functions to be best predicted by the network. Taken together, these results suggest that imputation methods solely relying on information encoded in the dataset, as is the case of scImpute, perform best when genes are captured across enough cells. However, for rarely captured genes, as is often the case for regulatory genes, making use of external information such as that encoded in transcriptional regulatory networks enables the imputation for more genes and improves the accuracy of the estimates.

Network-based imputation uncovers cluster markers and regulators

A popular application of scRNA-seq is the identification of discrete sub-populations of cells in a sample in order to, for example, identify new cell types. The clustering of cells and the visual 2D representation of single-cell data is affected by the choice of the dropout imputation method[12]. Therefore, we assessed the impact of dropout imputation on data visualization using t-Distributed Stochastic Neighbor Embedding (t-SNE)[13] on the hESC data before and after imputation by all four methods. The hESC dataset was particularly suitable in this case, because it was of high quality and it consisted of six well-annotated cell types. The t-SNE visualization was done on the complete dataset -- that is, with no masking procedure -- and on masked data (Fig. S7). We then asked how the clustering obtained with the complete dataset is recovered after imputing masked values. We observed that scImpute, the baseline and the network-based methods could recover the original cell clustering upon masking of up to 60% of quantified expression levels, while SAVER retained most of the noise in the data. At less stringent masking (35%), both the Network and Baseline - and SAVER, to a lesser extent - were able to recover the original cell clustering. scImpute imputation led to many fragmented clusters that were neither consistent with the clustering based on the full dataset, nor found after scImpute imputation without any masking, suggesting that these were not reflecting biologically relevant sub-types of cells.

We next asked to what extent the detection of cluster markers would be affected by the choice of the imputation method. Thus, we applied Seurat[14] to the hESC differentiation dataset, which was composed of a well-defined set of distinct cell types, before and after imputation. We then determined cluster markers as those genes that were differentially expressed between one cluster and all the others (Methods). We observed a strong overlap between markers detected before and after the tested imputation methods, exemplified in Fig. 3A for definitive endoderm cells (DECs). Spearman correlation between the obtained gene $\log_2$FC was also high (Fig. 3B), suggesting that the overall ranking of genes according to $\log_2$FC was maintained regardless of imputation.

The numbers of significant markers detected after network-based and baseline imputations were lower than for the other imputation methods (number of significant DEC markers at FDR < 0.05 and $|\log_2 FC| > 0.25$: Baseline = 1282, Network = 1740, Original = 3348, SAVER = 2244, scImpute = 5106). Imputation with scImpute led to the highest number of significant markers detected; however, these were consistently weakly enriched in biological functions when compared to the remaining methods and no imputation (Fig. S8), suggesting that the extra markers contained many false positives, which diluted biological signals. Conversely, network-based imputation tended to lead to markers enriched in cell type-specific processes (Fig. 3C).

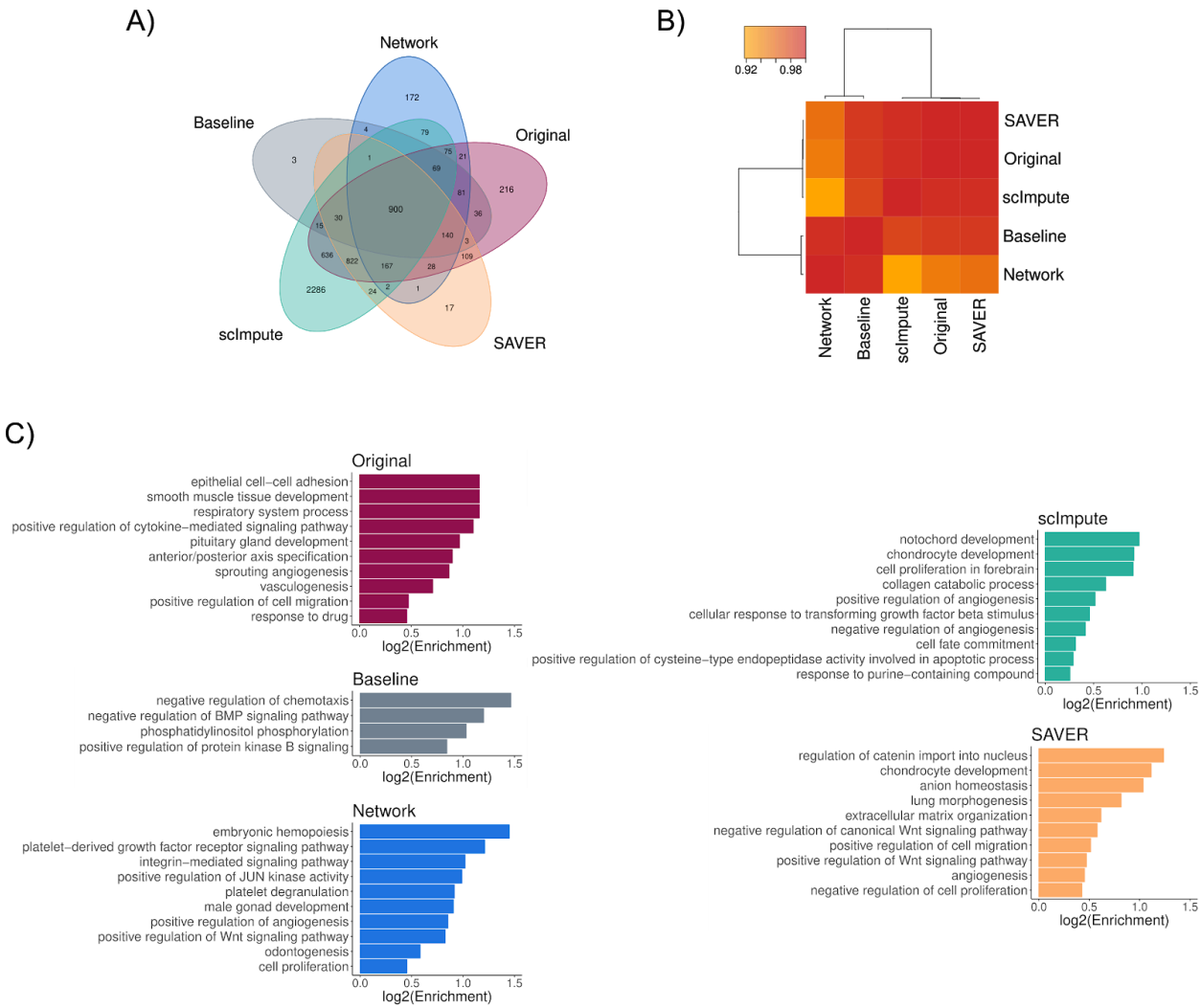

Figure 3: Detection of definitive endoderm cell (DEC)-specific markers before and after imputation. A) Overlap between significant (FDR < 0.05, $|\log_2 FC| > 0.25$) DEC markers detected with no dropout imputation (Original) and using the tested imputation methods. **B)** Spearman correlation matrix of detected $\log_2$FCs of all genes in DEC compared to the remaining clusters. **C)** Significantly enriched (p.value < 0.001) GO biological process terms among significant DEC markers detected with no dropout imputation and using the tested imputation methods.

Genes with regulatory functions are particularly important for understanding and explaining the transcriptional state of a cell. However, since genes with regulatory functions are often lowly expressed[15], they are frequently subject to dropouts. Since our analysis had shown that the network-based approach is especially helpful for lowly expressed genes (Fig. 2), we hypothesized that the imputation of transcript levels of regulatory genes would be particularly improved. In order to test this hypothesis, we further characterized those cluster markers that were exclusively detected using the network-based method. Indeed, we observed DNA-binding transcription factors to be enriched among those genes (Fig. 4A). Among these markers exclusively detected upon network-based imputation, the transcription factor EHF was the most significant trophoblast-specific. EHF is a known epithelium-specific transcription factor that has been described to control epithelial differentiation[16] and to be expressed in trophoblasts[17], even though at very low levels (EHF expression found among the first quintile of bulk TB RNA-seq data from the same authors). While EHF transcripts were not well captured in TB single-cell RNA-seq (only quantified in 39 out of 775 TB cells), a trophoblast-specific expression pattern was recovered after network-based imputation (Fig. 4B, upper panel), but not with any of the

other tested imputation methods (Fig. S8). Similarly, OSR1 has been described as a relevant fibroblast-specific transcription factor[18] which failed to be detected without imputation. Again, the network-based imputation was the only method that successfully recovered the fibroblast-specific expression pattern of OSR1 (Fig. 4B, lower panel), (Fig. S8). Interestingly, TWIST2 and PRRX1, described by Tomaru *et al.*[18] to interact with OSR1, also showed fibroblast-specific expression (Fig. S9). Taken together, these results suggest that imputation based on gene regulatory networks can recover the expression levels of relevant, lowly expressed regulators affected by dropouts.

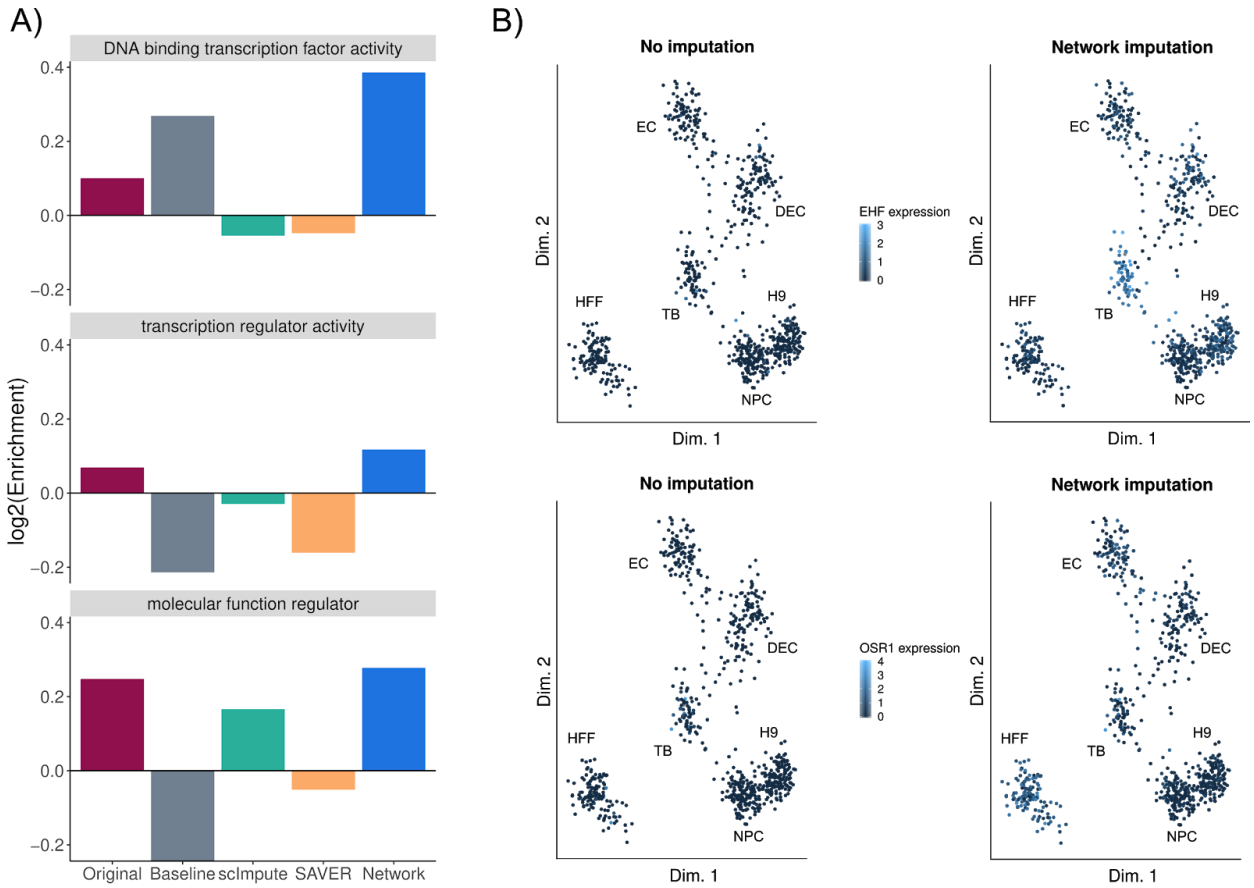

Figure 4: Detection of cell type-specific transcription factors is improved upon network-based imputation. A) Enrichment in selected GO molecular function terms related to transcriptional regulation among the genes uniquely detected after each imputation approach. Whereas genes detected with the Network imputation approach are most strongly enriched for genes with DNA-binding transcription factor activity, more generic terms like 'molecular function regulator' are less enriched. Hence, the Network approach seems particularly powerful in imputing the expression levels of lowly expressed regulatory genes. **B)** Projection of cells onto a low dimension representation of the data before imputation, using ZINB-WaVe[19]. Color represents expression levels of EHF (top) and OSR1 (bottom) before and after Network-based imputation. DEC: definitive endoderm cells; EC: endothelial cells; H9: undifferentiated human embryonic stem cells; HFF: human foreskin fibroblasts; NPC: neural progenitor cells; TB: trophoblast-like cells.

Discussion

A first important and surprising finding of our analysis is the fact that the sample-wide average expression performs well for the imputation of many genes (Table 1). As expected, genes whose expression levels were best imputed by this method were - compared to scImpute - characterized by lower variance across cells and by remaining undetected in relatively many cells (Fig. 2). A potential problem of methods based on co-clustering cells is that the number of observations per cluster can get very small, which makes the estimation of the true mean more unstable. Thus, the average using all cells is preferred when the gene was detected in only few cells and/or if the gene's expression does not vary much across cells. Further, our findings imply that cell-to-cell variation of gene expression is negligible for many genes.

Second, the consideration of external gene co-expression information for the dropout imputation substantially improved the performance in many cases, especially for lowly expressed genes. Since genes with regulatory functions are often lowly expressed[15], imputation of those genes might be critical for explaining expression variation between cells. The seemingly poor performance of SAVER can at least in part be attributed to our evaluation scheme, which is based on masking observed values. SAVER performs a re-scaling of predicted mean values (i.e. the means of the posterior distributions) that leads to larger deviations from the masked values. Another potential problem is the fact that SAVER learns gene-gene relationships from the scRNA-seq data itself, which may be imprecise or even impossible for genes with many dropouts. SAVER may very well maintain relative differences between genes and it clearly has the advantage of predicting the most likely true expression values also for observed genes (a feature not shared with any of the other methods evaluated in this work). Thus, one needs to chose the imputation method based on the specific goals.

The third -- and maybe most important -- conclusion is that the best performing imputation method is gene- and dataset-dependent. That is, there is no single best performing method. If the number of observations is high (many cells with detected expression) and if the expression quantification is sufficiently good, scImpute outperformed all other methods. Importantly, the technical quality of the quantification depends on the read counts, which in turn depends on gene expression, transcript length and mappability -- i.e. multiple factors beyond expression. If however, gene expression is low and/or too imprecise, scImpute was outcompeted by other methods. This finding led us to conclude that a combination of imputation methods would be optimal. Hence, we developed an R-package that determines 'on the fly' for each gene the best performing imputation method by masking observed values (i.e. *via* cross validation). This approach has the benefit that it self-adapts to the specificities of the dataset at hand. For example, the network-based approach might perform well in cell types where the assumptions of the co-expression model are fulfilled, whereas it might fail (for the same gene) in other cell types, where these assumptions are not met. Hence, the optimal imputation approach is gene- and dataset-dependent. An adaptive method selection better handles such situations. Another benefit of this approach is that the cross validation error can be used as a quantitative guide on how 'imputable' a given gene is in a specific scRNA-seq dataset. We have therefore implemented and tested this approach (see Supplementary Figure S11). The resulting R-package (called ADImpute) is open to the inclusion of future methods and it can be downloaded from http://cellnet.cecad.uni-koeln.de/adimpute.

We believe that this work presents a paradigm shift in the sense that we should no longer search for the single best imputation approach. Rather, the task for the future will be to find the best method for a particular combination of gene and experimental condition.

Methods

Pre-processing of cancer cell line data for gene regulatory network inference

Entrez IDs and corresponding gene symbols were retrieved from the NCBI (https://www.ncbi.nlm.nih.gov/gene/?term=human%5Borgn%5D). Genome annotation was obtained from Ensembl (*Biomart*). Finally, genes of biotype in protein coding, ncRNA, snoRNA, scRNA, snRNA were used for network inference. For CCLE[20], 768 cell lines that were used in Seifert et al.[8] were used. Raw CEL files were downloaded from https://portals.broadinstitute.org/ccle/ and processed using the R package RMA in combination with a BrainArray design file (HGU133Plus2_Hs_ENTREZG_21.0.0). Final expression values were in $\log_2$ scale. Expression levels and CNV data set from RNA-seq were downloaded from Klijn et al.[21]. Before combining, each dataset is $\log_2$ transformed and scaled to (0,1) for all genes in each sample using R function scale. Then datasets were merged and the function ComBat from the sva R package[22] was used to remove batch effect of the data source. The final combined data set contains 24641 genes in 1443 cell lines. Finally, expression levels of genes were subtracted by the average expression level across all cell lines of the corresponding gene.

Network inference based on stability selection

The network inference problem can be solved by inferring independent gene-specific sub-networks. We used the linear regression model from equation (1) to model the change in a target gene as dependent on the combination of the gene-specific CNA and changes in all other genes. Here the intercept is not included because the data is assumed to be centered. We used LASSO with stability selection[7] to find optimal model parameters α_{ij} .

The R package *stabs* was employed to implement stability selection and the *glmnet* package was used to fit the generalized linear model. Two parameters regarding error bounds were set with the cutoff value being 0.6 and the per-family error rate being 0.05. A set of stable variables were defined by LASSO in combination with stability selection. Then coefficients of the selected variables were estimated by fitting generalized linear models using the R function *glm*.

Network validation using TCGA data

Gene expression and gene copy number data of 14 different tumor cohorts (4548 tumor patients in total) from TCGA collected in a previous study[8] were used for validation. We examined the predictive power of our inferred networks on each TCGA cohort by predicting the expression level of each gene for each tumor using the corresponding copy number and gene expression data. The quality of the prediction for each TCGA cohort was quantified as the correlation between the original gene expression levels in the TCGA dataset and the corresponding expression levels predicted by the provided network for each gene across all patients in a

cohort (Fig. S1). A strong positive correlation for a gene suggests high predictive power by the network on the respective gene.

Single-cell test data processing

Human embryonic stem cell differentiation data[9] were downloaded from GEO (accession number GSE75748) in the format of expected counts. Only snapshot single-cell data were used in this work (file GSE75748_sc_cell_type_ec.csv.gz). The downloaded data were converted to TPM (transcripts per million). Gene length was determined as the average length of all transcripts (provided by the *biomaRt* package) for a given gene. Oligodendroglioma data[10] were downloaded from GEO (accession number GSE70630) as $\log_2$(TPM/10+1) and converted back to TPM. Healthy human pancreas data[11] were downloaded from GEO (accession number GSE84133) as UMI counts and converted to TPM as described.

Dropout imputation

Version 0.0.9 of scImpute[2] was used for dropout imputation, in “TPM” mode, without specifying cell type labels. The parameters were left as default, except for drop_thre = 0.3, as the default of 0.5 resulted in no imputations performed. Cell cluster number (Kcluster) was left at the default value of 2 for imputation of the oligodendroglioma and healthy pancreata datasets and set to 6 for the hESC differentiation dataset, in order to match the number of cell clusters identified by the authors[9]. For all other imputation methods, the data was converted to $\log_2$(TPM+1). SAVER 1.1.1 was used with size.factors = 1. For dropout imputation by average expression (‘baseline’), gene expression levels were converted to $\log_2$(TPM+1) and the average expression of each gene across all cells, excluding zeros, was used for imputation. For network-based imputation, expression values were converted to $\log_2$(TPM+1) and centered gene-wise across all cells. The original centers were stored for posterior re-conversion. Subsequently, cell-specific deviations of expression levels from those centers were predicted using either equation (5) or the following iterative procedure. During the iteration genes were first predicted using all measured predictors. Subsequently, genes with dropout predictors were re-predicted using the imputed values from the previous iteration. This was repeated for at most 50 iterations. The obtained values were added to the gene-wise centers. We note that, while the values after imputation cannot be interpreted as TPMs, as the sum of the expression levels per sample is no longer guaranteed to be the same across samples. However, one could still perform a new normalization by total signal (sum over all genes) to overcome this issue.

Imputation error analysis

In order to compare the imputation error of the tested methods, we randomly masked (set to zero) a fraction of the values for each gene (Fig. 1) - 35% for the hESC differentiation dataset, 10% for the oligodendroglioma dataset and 8% for both healthy pancreas datasets - or cell (Fig. S1) - 30%. This unbiased masking scheme is in agreement with previous work[23]. The differing percentages of masked values per gene in each dataset result in a comparable sparsity of the data (around 84% missing values) after masking. Imputation was performed with each of the four tested methods separately and the imputed masked entries were then compared to the

original ones. As dropout-specific performance measure, we used the squared imputation error corrected for average gene expression (log-transformed TPM):

$$\frac{(original - imputed)^2}{avg\ expression + 0.1} \qquad (6)$$

The weighting prevents that the average error is dominated by highly expressed genes, i.e. some weighting is necessary. Another alternative would have been to just divide by the average expression (without adding 0.1). In that case however, the error estimates would have been heavily dominated by very lowly expressed genes since their average is very close to zero. Thus, the weighting proposed here is a fair compromise considering both, highly and lowly expressed genes.

Dimensionality reduction and marker detection

Dimensionality reduction on the original hESC differentiation data (Fig. 4B) was performed using ZINB-WaVe, implemented in the R package *zinbwave* [19]. H1 and TB cells in Batch 3 were removed to avoid confounding batch effects and, for the remaining cells, 2 latent variables were extracted from the information contained in the top 1000 genes with highest variance across cells. Batch information and the default intercepts were included in the ZINB-WaVe model, using *epsilon* = 1000. K-means clustering (k = 6) on the 2 latent variables strongly matched the annotated cell type labels (0.977 accuracy), confirming the reliability of this approach. t-SNE was performed on the hESC differentiation data ($\log_2$(TPM+1)) before and after imputation (Fig. S7) using the *Rtsne* R package. Cluster-specific markers were detected from the $\log_2$(TPM+1) data using Seurat[14]. Detection rate was regressed out using the *ScaleData* function with *vars.to.regress* = *nGene*. Markers were detected with the *FindAllMarkers* function, using MAST[24] test and setting *logfc.threshold* and *min.pct* to 0, and *min.cells.gene* to 1.

GO term enrichment and transcription factor analyses

All GO term enrichment analyses were performed with the *topGO* R package[25], using the *weight01* algorithm for GO hierarchy handling, with Fisher test. To test for enrichment on GO molecular functions among genes best predicted by each imputation method (Fig. S6), all the genes for which the imputation error of at least one method could be computed were used as background. Enrichment in GO biological process terms among cluster-specific markers (Fig. 3C) was performed using as foreground the set of markers detected by Seurat with FDR < 0.05 and |logFC| > 0.25, and as background all genes in the Seurat result (both significant and non-significant). FDR correction of the enrichment results was done using the *p.adjust* function in R. $\log_2$ enrichment was quantified as the $\log_2$ of the ratio between the number of significant and expected genes in each term. For regulatory GO molecular function term enrichment analyses (Fig. 4A), significant (FDR < 0.05 and |logFC| > 0.25) markers uniquely detected without / with each imputation method were combined across all clusters and tested for enrichment in the terms “DNA-binding transcription factor activity”, “transcription regulator activity” and “molecular function regulator” against the background of all genes obtained as the result of Seurat (both significant and non-significant).

To identify transcription factors (TFs) among cluster markers exclusively detected using the network-based method, a curated TF list was downloaded from http://www.tfcheckpoint.org/index.php/browse.

Determination of the optimal imputation method per gene

In order to determine the best performing imputation method for each gene, 70% of the cells in each dataset were used as training data, where a percentage of the expression values were masked, as previously described. The remaining 30% were used for testing. After masking, each of the three imputation methods was applied to the training data and the imputed values of masked entries were then compared to the measured values. The weighted Mean Squared Error (MSE) was computed for each gene with masked entries:

$$\frac{avg((original - imputed)^2)}{avg\ expression + 0.1} \qquad (7)$$

For each gene, the method leading to the smallest MSE was chosen as optimal.

Acknowledgements

The results here shown are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. A.C.L. received support by the Cologne Graduate School of Ageing Research. X.W. received financial support from the National Natural Science Foundation of China (61871463) and Natural Science Foundation of Fujian Province of China (2017J01068).

We gratefully acknowledge help from Dr. Michael Seifert (TU Dresden, Germany) on the construction of the transcriptional regulatory network.

Author Contribution

AB envisioned the study. XW implemented and performed the network training and testing. ACL implemented and tested the dropout imputation method. All authors contributed to the writing of the manuscript.

References

1. van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716–729.e27 (2018).
2. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997 (2018).
3. Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N. & Garry, D. J. DrImpute: imputing

dropout events in single cell RNA sequencing data. *BMC Bioinformatics* **19**, 220 (2018).

4. Moussa, M. & Mandoiu, I. Locality Sensitive Imputation for Single-Cell RNA-Seq Data. (2018). doi:10.1101/291807
5. Huang, M. *et al.* SAVER: Gene expression recovery for UMI-based single cell RNA sequencing. (2017). doi:10.1101/138677
6. Meinshausen, N. & Yu, B. *Lasso-type recovery of sparse representations for high-dimensional data*. (2006). doi:10.21236/ada472998
7. Meinshausen, N. & Bühlmann, P. Stability selection. *J. R. Stat. Soc. Series B Stat. Methodol.* **72**, 417–473 (2010).
8. Seifert, M., Friedrich, B. & Beyer, A. Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis. *Genome Biol.* **17**, 204 (2016).
9. Chu, L.-F. *et al.* Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* **17**, 173 (2016).
10. Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
11. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems* **3**, 346–360.e4 (2016).
12. Zhang, L. & Zhang, S. Comparison of computational methods for imputing single-cell RNA-sequencing data. (2017). doi:10.1101/241190
13. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
14. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
15. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).

16. Tugores, A. *et al.* The epithelium-specific ETS protein EHF/ESE-3 is a context-dependent transcriptional repressor downstream of MAPK signaling cascades. *J. Biol. Chem.* **276**, 20397–20406 (2001).

17. Boyd, C. A. R. Review: Epithelial aspects of human placental trophoblast. *Placenta* **34 Suppl**, S24–6 (2013).

18. Tomaru, Y. *et al.* A transient disruption of fibroblastic transcriptional regulatory network facilitates trans-differentiation. *Nucleic Acids Res.* **42**, 8905–8913 (2014).

19. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 284 (2018).

20. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

21. Klijn, C. *et al.* A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.* **33**, 306–312 (2015).

22. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics* **3**, e161 (2007).

23. Talwar, D., Mongia, A., Sengupta, D. & Majumdar, A. AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Sci. Rep.* **8**, 16329 (2018).

24. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16**, (2015).

25. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).