False signals induced by single-cell imputation [version 1; referees: awaiting peer review]

Tallulah S. Andrews, Martin Hemberg
Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK

Abstract

**Background:** Single-cell RNASeq is a powerful tool for measuring gene expression at the resolution of individual cells. A significant challenge in the analysis of this data is the large amount of zero values, representing either missing data or no expression. Several imputation approaches have been proposed to deal with this issue, but since these methods generally rely on structure inherent to the dataset under consideration they may not provide any additional information.

**Methods:** We evaluated the risk of generating false positive or irreproducible results when imputing data with five different methods. We applied each method to a variety of simulated datasets as well as to permuted real single-cell RNASeq datasets and consider the number of false positive gene-gene correlations and differentially expressed genes. Using matched 10X Chromium and Smartsseq2 data from the Tabula Muris database we examined the reproducibility of markers before and after imputation.

**Results:** The extent of false-positive signals introduced by imputation varied considerably by method. Data smoothing based methods, MAGIC and knn-smooth, generated a very high number of false-positives in both real and simulated data. Model-based imputation methods typically generated fewer false-positives but this varied greatly depending on how well datasets conformed to the underlying model. Furthermore, only SAVER exhibited reproducibility comparable to unimputed data across matched data.

**Conclusions:** Imputation of single-cell RNASeq data introduces circularity that can generate false-positive results. Thus, statistical tests applied to imputed data should be treated with care. Additional filtering by effect size can reduce but not fully eliminate these effects. Of the methods we considered, SAVER was the least likely to generate false or irreproducible results, thus should be favoured over alternatives if imputation is necessary.

**Keywords**
Gene expression, single-cell, RNA-seq, Imputation, Type 1 errors, Reproducibility

This article is included in the Bioconductor gateway.
Introduction

Single-cell RNASeq (scRNASeq) is a powerful technique for assaying the whole transcriptome at the resolution of individual cells. Although experimental protocols have evolved rapidly, there is still no strong consensus on how to best analyse the data. An important challenge to analysing scRNASeq data is the high frequency of zero values, often referred to as dropouts, and the overall high levels of noise due to the low amounts of input RNA obtained from individual cells. Recently there have been four methods published (Gong et al., 2018; Huang et al., 2018; Li & Li, 2018; van Dijk et al., 2018) which attempt to address these challenges though imputation, with several more under development (Deng et al., 2018; Mongia et al., 2018; Moussa & Mandoiu, 2018; Wagner et al., 2017).

Imputation is a common approach when dealing with sparse genomics data. A notable example has been the improvements to GWAS sensitivity and resolution when using haplotype information to impute unobserved SNPs (Visscher et al., 2017). Unlike scRNASeq data, this imputation employs an external reference dataset, often the 1000 Genomes project, to infer the missing values (Chou et al., 2016). Such a reference does not yet exist for scRNASeq data, and thus imputation methods can only use information internal to the dataset to be imputed. As a result there is a degree of circularity introduced into the dataset following imputation which could result in false positive results. Zero values in scRNASeq may arise due to low experimental sensitivity, e.g. sequencing sampling noise, technical dropouts during library preparation, or because biologically the gene is not expressed in the particular cell. Thus, one challenge when imputing expression values is to distinguish true zeros from missing values.

Many imputation methods, such as SAVER (Huang et al., 2018), DrImpute (Gong et al., 2018) and sclImpute (Li & Li, 2018), use models of the expected gene expression distribution to distinguish true biological zeros from zeros originating from technical noise. Because these gene expression distributions assume homogenous cell populations, they first identify clusters of similar cells to which an appropriate mixture model is fitted. Values falling above a given probability threshold to originate from technical effects are subsequently imputed. For example, sclImpute models log-normalized expression values as a mixture of gamma-distributed dropouts and normally-distributed true observations. Alternatively some scRNASeq imputation methods perform data smoothing. In contrast to imputation, which only attempt to infer values of missing data, smoothing reduces noise present in observed values by using information from neighbouring data points. Both MAGIC (van Dijk et al., 2018) and knn-smooth (Wagner et al., 2017) perform data smoothing for single-cell data using each cell’s k nearest neighbours either through the application of diffusion models or weighted sums respectively.

Previous benchmarking of these imputation methods was based on positive controls, i.e. the ability to recover true signals within noisy data (Zhang & Zhang, 2018); the potential for false signals to be introduced into a dataset by these imputation methods was not considered, and it was concluded that most imputation methods provide a small improvement. We consider negative controls to evaluate the risks of introducing false positive when using imputation for single-cell datasets. Testing of the four published imputation methods, MAGIC, SAVER, sclImpute, and DrImpute and one currently unpublished method, knn-smooth, revealed that all methods can introduce false positive signals into data. While some methods, performed well on simulated data, permuting real scRNASeq data revealed high variability in performance on different datasets. We show that statistical tests applied to imputed data should be treated with care, and that results found in imputed data may not be reproducible.

Methods

Five different single-cell RNASeq imputation methods were tested: SAVER (Huang et al., 2018), DrImpute (Gong et al., 2018), sclImpute (Li & Li, 2018), MAGIC (van Dijk et al., 2018) and knn-smooth (Wagner et al., 2017). Unless specified otherwise these were run with default parameters (Table 1). Each method was applied to either the raw-counts or log2 counts per million normalized data, as calculated scater (McCarthy et al., 2017), as appropriate.

Negative binomial simulations

As an initial test of imputation methods and to understand the effect of various method-specific parameters on imputation we simulated data from a negative binomial model. Expression matrices containing 1000 cells, equally spread across two cell-types, and 500 genes, with mean expression ranging from \(10^{-5}\)–\(10^4\), were simulated. Half of the genes were differentially expressed (DE) by an order of magnitude between the two cell-types, half were drawn independently. Ten such expression matrices were independently simulated. Each imputation method was run on each replicate with a range of parameter values (Table 1). Signiﬁcant gene-gene correlations were identiﬁed using Spearman correlation with a conservative Bonferroni multiple testing correction to avoid distributional assumptions on the imputed values. Correlations involving not DE genes or in the incorrect direction were considered false positives.

Splatter simulations

Splatter (Zappia et al., 2017) was used to generate 60 simulated single-cell RNASeq count matrices using different combinations of parameters (Table 2). Each simulated dataset contained 1,000 cells split into 2–10 groups and 1,000–5,000 genes of which 1–30% were differentially expressed across the groups. For simplicity all groups were equally sized and were equally different from one another. Half the simulations assumed discrete differentiated groups, whereas the other half used the continuous differentiation path model. We also considered the effect of four different amounts of added dropouts plus the no-added dropout model.

Accuracy of each imputation method was evaluated by testing for differentially expressed (DE) genes between the groups used to simulate the data. To avoid issues of different imputation methods resulting in data best approximated by different probability distribution, we employed the non-parametric Kruskal-Wallis test (Kruskal & Wallis, 1952) with a 5% FDR to identify significant DE genes. Since this test is relatively
Table 1. Imputation methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Date</th>
<th>Parameter(s)</th>
<th>Range</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>scImpute</td>
<td>2018</td>
<td>Dropout threshold, Number of clusters</td>
<td>0-1 (default: 0.5), Correct value given the simulation</td>
<td>(Li &amp; Li, 2018)</td>
</tr>
<tr>
<td>DrImpute</td>
<td>2018</td>
<td>Remaining zeros, Number of clusters</td>
<td>0-1 (default: 0), Correct value given the simulation</td>
<td>(Gong et al., 2018)</td>
</tr>
<tr>
<td>SAVER</td>
<td>2018</td>
<td>Which genes to impute</td>
<td>Top 1%-100% most highly expressed (default: 100%)</td>
<td>(Huang et al., 2018)</td>
</tr>
<tr>
<td>MAGIC</td>
<td>2018</td>
<td>Diffusion time, K neighbours</td>
<td>1–8 (default: allow algorithm to choose), 5–100 (default: 12)</td>
<td>(van Dijk et al., 2018)</td>
</tr>
<tr>
<td>knn smooth</td>
<td>2017</td>
<td>K neighbours</td>
<td>5–100 (default: number of cells / 20)</td>
<td>(Wagner et al., 2017)</td>
</tr>
</tbody>
</table>

Table 2. Splatter parameters.

<table>
<thead>
<tr>
<th>nGenes*</th>
<th>%DE (total)*</th>
<th>Dropouts (midpoint)</th>
<th>nGroups</th>
<th>Method</th>
<th>Seed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>1%</td>
<td>None</td>
<td>2</td>
<td>Groups</td>
<td>8298</td>
</tr>
<tr>
<td>2000</td>
<td>10%</td>
<td>1</td>
<td>5</td>
<td>Paths</td>
<td>2900</td>
</tr>
<tr>
<td>5000</td>
<td>30%</td>
<td>3</td>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Randomly selected for each possible combination of the other four parameters.

low-power it is likely to underestimate the number of DE genes compared to alternatives. When filtering DE genes by effect size, in addition to significance, we used the maximum log2-fold-change across all pairs of clusters.

Permutated Tabula Muris datasets
Six 10X Chromium and 12 Smartseq2 datasets were chosen from the Tabula Muris (The Tabula Muris Consortium et al., 2017) consortium data such that: i) there were at least two cell types containing >5% of the total cells and ii) there were between 500–5,000 cells after filtering (Table S1). Each dataset was preprocessed to remove cell-types accounting for <5% of total cells, and any cells not assigned to a named cell-type. Genes were filtered to remove those detected in fewer than 5% of cells.

We selected the two most similar cell-types in each dataset using the Euclidean distance between their mean expression profiles. Differential expression of each gene between these cell-types was evaluated using a Mann-Whitney-U test on the log2-normalized counts. Genes with a raw p-value > 0.2 were then permuted across the selected cell-types to eliminate any residual biological signals. Permuted raw counts were obtained by de-logging and de-normalizing the permuted log2-normalized expression to avoid library-size confounders.

Each imputation method was applied to the full dataset after permutation using default parameters (Table 1). False-positives introduced by each imputation was assessed by applying the Mann-Whitney-U test to test for differential expression between the two chosen cell-types. A Bonferroni multiple-testing correction was applied to ensure a consistent level of expected total false positives of less than 1.

Reproducibility of markers
We utilized the six tissues for which there exists matching Smart-seq2 and 10X Chromium data from the Tabula Muris (The Tabula Muris Consortium et al., 2017) to evaluate the reproducibility of imputation results. These datasets were filtered as described above, and any cell-types not present in both pairs of the matching datasets were excluded. Each imputation method was applied to the datasets without any permutation.

Marker genes were identified in each imputed dataset using a Mann-Whitney-U test to compare each cell-type against all others, and effect size was calculated as the area under the ROC curve for predicting each cell-type from the others (Kiselev et al., 2017). Genes were assigned to the cell-type for which they had the highest AUC. Significant marker genes were defined for each imputed dataset using a 5% FDR and an AUC over a particular threshold. Reproducibility was evaluated by determining the number of genes that were significant markers in both of a matching pair of datasets and were markers of the same cell-type.

Results
We tested three published imputation methods, SAVER (Huang et al., 2018), scImpute (Li & Li, 2018) and DrImpute...
(Gong et al., 2018), as well as two data-smoothing methods MAGIC (van Dijk et al., 2018) and knn-smooth (Wagner et al., 2017). We applied each method with the default parameter values (Table 1) to data simulated from a simple negative binomial, since technical noise in scRNASeq data has been observed to follow a negative binomial distribution (Grün et al., 2014). Half the simulated genes were differentially expressed (DE), thus highly correlated with each other, the rest were drawn completely independently. These simulations did not include different library-sizes, batch effects, or zero-inflation to eliminate all possible sources of false-signals that imputation method might mistake for true biology. Thus, these simulations represent the simplest most straightforward case with no technical confounders. Only SAVER strengthened the correlations between lowly expressed DE genes without generating false positive gene-gene correlations between independently drawn genes (Figure 1A). Since SAVER models expression data using a negative binomial, it is expected to perform well on this simulated data. MAGIC generated very strong false positive correlations \((r > 0.75)\) at all expression levels, whereas DrImpute, which only imputes zero values, created false positive correlations mostly among lowly expressed genes. Knn-smooth and scImpute produced a few false-positive correlations among moderately-expressed genes using default parameters.

Choice of parameter values has a large influence imputation results (Figure 1B). Four of the imputation methods required the user to set at least one parameter \(a \text{ priori}\), only SAVER did not. We varied the thresholds scImpute and DrImpute use to determine which zeros to impute. For scImpute some of the lower and moderate expression values were imputed even at a very strict probability threshold \((p > 0.8)\), but changing the threshold had little effect on the imputation. As expected for DrImpute, imputing a greater proportion of zeros generated more false

![Figure 1](https://example.com/figure1.png)

**Figure 1.** False gene-gene correlations induced by single-cell imputation methods. (A) Gene-gene correlations before and after imputation using suggested parameter values: SAVER (all genes), MAGIC \((k=12, t=3)\), knn \((k=50)\), scImpute \((\text{threshold}=0.5)\), DrImpute \((\text{remaining zeros}=0)\). Coloured bars indicate genes highly expressed (red) or lowly expressed (blue) in one cell population vs the other, or genes not differentially expressed between the populations (grey). Genes are ordered left to right by DE direction then by expression level (high to low). (B) False positive gene-gene correlations as imputation parameters are changed. Dashed lines are 95% CIs based on 10 replicates. See Figure S1 for true positive rate of gene-gene correlations across the same parameters.
positives. Knn-smooth and MAGIC both perform data smoothing using a k-nearest-neighbour graphs between cells. Increasing the number of nearest-neighbours ($k$) produces smoother data and more false-positive correlations (Figure 1B). MAGIC provides a default value for $k$ but no indication of how this parameter should be adjusted for different sized datasets, whereas knn smooth provided no default value but a rough suggestion to scale the value depending on the total number of cells. MAGIC also utilizes a second parameter, time ($t$), for the diffusion process acting on the graph which by default is algorithmically estimated for the dataset. Longer diffusion times produce smoother data and more false positives.

These simple simulations contained only two cell-types and no technical confounders such as library-size or inflated dropout rates that are observed in some scRNASeq datasets. For a more comprehensive evaluation of imputation methods we simulated data using Splatter (Zappia et al., 2017). We simulated data with 1,000 cells split into 2–10 groups and 1,000–5,000 genes of which 1–30% were differentially expressed across the groups. We considered four different levels of zero inflation and no zero inflation (Table 2). Each simulated dataset was imputed with each method using the default parameters (Table 1). To score each imputation we considered the accuracy of identifying differentially expressed genes between the groups using the non-parametric Kruskal-Wallis test (Kruskal & Wallis, 1952).

None of the imputation methods significantly outperformed the others or the unimputed data based on the sensitivity and specificity. While both knn-smooth and MAGIC have increased sensitivity they have very low specificity, whereas SAVER and scImpute are very similar to the un-imputed data with high specificity but relatively low sensitivity (Figure 2A & B). DrImpute was in between the two extremes with somewhat higher sensitivity and lower specificity than SAVER and scImpute. Both scImpute and DrImpute are designed specifically to only impute excess zeros but neither showed a clear improvement over the raw counts when the simulations contained various levels of zero inflation.

![Figure 2](image-url)  
Figure 2. Accuracy of detecting differentially expressed (DE) genes in splatter simulations before and after imputation with each method. (A & B) Different imputation methods choose a different trade-off between sensitivity and specificity. (C) Zero inflation decreases sensitivity of DE which most imputation methods fail to correct. (D) Strong true signals (high proportion of genes DE) decreases specificity particularly for data-smoothing methods.
of zero inflation (Figure 2C). However, all methods except SAVER readily introduced false-positive signals, as demonstrated by a drop in specificity, when 30% of genes were DE (Figure 2D). We hypothesize that slight biases due to library-size normalization in the presence of strong biological differences, may be amplified by the imputation methods since we also observe a significant but smaller drop in specificity for the normalized but un-imputed data. Biases due to counts-per-million library-size normalization in the presence of strong DE are a known issue from bulk RNASeq analysis (Bullard et al., 2010).

It’s possible that the bulk of false-positives generated by imputation methods result from small biases or sampling noise being amplified to reach statistical significance. If this is true, then filtering DE genes by magnitude in addition to significance should restore the specificity of such tests on imputed data. We observed this to be the case when an additional threshold was set based on the Xth percentile highest log2 fold-change across the whole dataset (Figure 3). However, sensitivity also declined as the fold-change threshold was made more stringent. This suggests the fundamental trade-off between sensitivity and specificity cannot be overcome by imputation.

Splatter is a widely used simulation framework for scRNASeq but may not fully capture the complexities of real scRNASeq data. To test the performance of each imputation method on real scRNASeq data we selected 12 tissues from the Tabula Muris database (The Tabula Muris Consortium et al., 2017) and applied the imputation methods to the Smartseq2 and 10X data separately. Since the ground truth is not known for these data, we selected two cell-types from each dataset and permuted the expression of those genes that were not differentially expressed between them (p > 0.2) to generate a set of genes that we could confidently consider as being not differentially expressed (Methods). Using these as ground truth we could estimate the number of false positive differentially expressed genes introduced by each imputation method. Strikingly, we observed a very high variability between datasets which appears to be unrelated to the experimental platform (Figure 4A & B). MAGIC, scImpute and knn-smooth consistently produced large numbers of false positives (40–80%). Whereas, DrImpute and SAVER
were extremely variable producing few to no false positives in some datasets and over 90% false positives in others.

To complement the false positives in the permuted data, we considered whether imputed signals in the original datasets were reproducible in both 10X and Smart-seq2 data. We identified marker genes using a Mann-Whitney-U test, comparing one cell-type to the others in that tissue. Markers were filtered by significance (5% FDR) and magnitude (AUC > 0.75). Each marker was assigned to the cell-type for which it had the highest AUC. Reproducibility was measured as the number of markers that were significant in both datasets and the proportion of those that were markers for the same cell-type. All of the imputation methods increased the number of significant markers (Figure 4C). However, many of these were assigned to contradictory cell-types. Without imputation, 95% of genes that were significant markers in both datasets were highly expressed in the same cell-type. After imputation, this dropped to only 80–90%, except when SAVER was used for the imputation, suggesting that many of the imputed markers are incorrect. Decreasing the magnitude threshold of the markers leads to even more contradictory results in the imputed datasets (Figure 5). While unimputed data retained >90% concordance in cell-type assignments of significant markers regardless of the AUC threshold, this fell to 60–80% in imputed data when a low AUC threshold is used.

The imputation methods produced different distortions of the gene expression values (Figure 6). When applied to the permuted pancreas data, SAVER made only slight adjustments to the gene expression values. scImpute compressed the gene expression distributions into a more gaussian shape. DrImpute shifted zero values up into the higher mode of the distribution if present. In contrast, MAGIC and knn-smooth tended to generate bimodal expression distributions. The tendency towards bimodality could be problematic for downstream analysis since many methods, e.g. PCA and differential expression, assume either negative binomial or gaussian distributions. Many of these genes were differentially expressed after imputation, despite being permuted previously. Interestingly, the direction of differential expression was not always consistent across imputation methods, for instance Zfp606 was more highly expressed in PP cells than A cells after imputation using MAGIC but the inverse was true after imputing with knn-smooth.

**Discussion**

We have shown that imputation for scRNASeq data may introduce false-positive results when no signal is present. On simulated data all the methods except SAVER generated some degree of false positives (Figure 1 & Figure 2). We find a fundamental trade-off between sensitivity and specificity which imputation cannot overcome (Figure 2 & Figure 3). On permuted...
**Figure 5.** Filtering by the magnitude of expression differences restores specificity in imputed data. Markers were identified in 10X Chromium and Smartseq2 data for mouse muscle. The number of markers (bars, left axis) and proportion reproducible across both datasets (line, right axis) are plotted. Only significant markers (5% FDR) exceeding the AUC threshold were considered.

**Figure 6.** Examples of false positive DE induced by imputation of Pancreas Smartseq2 data. Unimputed indicates the permuted normalized log-transformed expression. Red = PP cell, Blue = A cell. * = p < 0.05, ** = significant after Bonferroni correction.
real data, imputation results were more variable (Figure 4), and even SAVER generated large numbers of false positives in some datasets. Imputation also reduced the reproducibility of marker genes, unless strict magnitude thresholds were imposed (Figure 4 & Figure 5). In addition to false-positives, distortions in expression distributions (Figure 6) may cause imputed data to violate assumptions of some statistical tests.

We found a trade-off between sensitivity and specificity across methods (Figure 2). MAGIC and knn-smooth which are data-smoothing methods, as such they adjust all expression values not just zeros. Since they impose larger alterations on the data, these methods generate many more false positives than methods which only impute zero values. However, they also have a greater sensitivity. In contrast, model-based methods which only impute low expression values, generated fewer false positives but had minimal improvements to sensitivity.

This trade-off between sensitivity and specificity also emerges if one employs an effect size threshold to reduce false-positives generated by imputation (Figure 3 & Figure 5). While using a strict effect size threshold can recover a reasonable specificity for the data-smoothing methods, doing so eliminates the improvements to sensitivity. Likewise, adding an effect size threshold can preserve reproducibility of imputation results but doing so largely eliminates the benefit in terms of number of markers identified.

These trade-offs reflect the fundamental limitation of single-cell RNASeq imputation, namely that it can only use the information present in the original data. While imputation in other fields often uses external references or relationships for the imputation, scRNASeq imputation only draws on structure within the dataset itself. Hence no new information is gained, making it analogous to simply lowering the significance threshold of any statistical test applied to the data (Fawcett, 2006). However, imputation based on external reference datasets may become possible as various cell-type atlases are completed; however these will be limited to those species and tissues that have been systematically catalogued (Han et al., 2018; Rozenblatt-Rosen et al., 2017; The Tabula Muris Consortium et al., 2017; Zeisel et al., 2018). Alternatively, models could be developed to use gene-gene correlations derived from large external databases of expression data (Obayashi et al., 2008), while more generalizable such methods may not capture cell-type specific relationships.

Of the methods we tested, SAVER was the least likely to generate false-positives, but it’s performance depended on how well data conformed to the negative binomial model it is based on. If imputation is used, combining SAVER with an effect size threshold is the best option to avoid irreproducible results. Alternatively, verifying the reproducibility of results across multiple datasets or multiple imputation methods can eliminate some false positives. However, our results highlight that statistical tests applied to imputed data should be treated with care. Although a previous benchmarking study showed good results for positive controls, our study highlights the importance of considering negative controls when evaluating imputation methods.

**Data and software availability**

**Tabula Muris data**


**R packages**

MAGIC: Rmagic (v0.1.0) https://github.com/KrishnaswamyLab/MAGIC

DrImpute: DrImpute (v1.0) https://github.com/ikwak2/DrImpute

scImpute: scImpute(v0.0.8) https://github.com/Vivianstats/scImpute

SAVER: SAVER(v1.0.0) https://github.com/mohuangx/SAVER


Scater: scater(v1.6.3) : https://www.bioconductor.org/packages/release/bioc/html/scater.html

Splatter: splatter(v1.2.2) : https://bioconductor.org/packages/release/bioc/html/splatter.html

Permute: permute(v0.9-4) : https://cran.r-project.org/web/packages/permute/index.html

**Grant information**

Funding was provided by the Wellcome Trust Sanger Institute Core Funding.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Supplementary material**

Supplementary File 1: File containing Table S1 (Tabula Muris permuted datasets) and Figure S1 (Sensitivity single-cell imputation methods).

Click here to access the data
References


The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com