bioRxiv

ARIC: Accurate and robust inference of cell type proportions from bulk gene expression or DNA methylation data

Wei Zhang, Hanwen Xu, Rong Qiao, Bixi Zhong, Xianglin Zhang, Jin Gu, Xuegong Zhang, Lei Wei* and Xiaowo Wang*

Ministry of Education Key Laboratory of Bioinformatics; Center for Synthetic and Systems Biology; Bioinformatics Division, Beijing National Research Center for Information Science and Technology; Department of Automation, Tsinghua University, Beijing, 100084, China

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Quantifying the cell proportions, especially for rare cell types in some scenarios, is of great value to track signals related to certain phenotypes or diseases. Although some methods have been pro-posed to infer cell proportions from multi-component bulk data, they are substantially less effective for estimating rare cell type proportions since they are highly sensitive against feature outliers and collinearity. Here we proposed a new deconvolution algorithm named ARIC to estimate cell type proportions from bulk gene expression or DNA methylation data. ARIC utilizes a novel two-step marker selection strategy, including component-wise condition number-based feature collinearity elimination and adaptive outlier markers removal. This strategy can systematically obtain effective markers that ensure a robust and precise weighted u-support vector regression-based proportion prediction. We showed that ARIC can estimate fractions accurately in both DNA methylation and gene expression data from different experiments. Taken together, ARIC is a promising tool to solve the deconvolution problem of bulk data where rare components are of vital importance.

Introduction

High-throughput biological technologies, such as microarrays, RNA-seq and whole-genome bisulfite sequencing, provide us informative approaches to investigate various biological samples collected from laboratories or clinical trials [1, 2]. However, plenty of these biological samples are complex mixtures of many different cell types without knowing their accurate proportions. Meanwhile, almost all physiological and pathological processes in multicellular organisms involve multiple cell types, each playing its particular roles [3]. Therefore, estimating the proportions of all or some specific cell types from bulk high-throughput data helps to understand the mechanism of biological processes and decouple signals related to certain phenotypes or diseases. For instance, the proportion of white blood cells can indicate the severity of immunological rejection after transplanting kidneys [4]. Quantifying the density of tumor infiltrating lymphocytes (TILs) like $CD3^+$ and $CD8^+$ T cells helps to predict patient survival in kinds of cancers [5-9]. Estimating placental signals in plasma cell-free DNA (cfDNA) of pregnant women can characterize the development of fetuses [10]. Detecting tumor-derived DNA fragments from plasma cfDNA can help us reveal the origin and development of cancers [11].

The fractions of minority cell types are of special interest in some deconvolution tasks. For example, the fraction of tumor-derived cfDNA in total plasma cfDNA, which is ultra-low for patients with early-stage cancer, is of vital importance for tumor detection [12]. TILs, which are promising biomarkers for clinical outcome prediction, also exhibit low fractions in many cancer tissues [5, 7, 13, 14]. Precise estimation of the fractions of certain cell types, especially rare cell types for some scenarios, could benefit a lot to discover signatures related to certain phenotypes or diseases.

Currently, many cell-type-specific omics data such as DNA methylation and gene expression profiles have been produced and can be accessed from various resources, like The Cancer Genome Atlas (TCGA) program [15], Encyclopedia of DNA Elements (ENCODE) project [16] and Gene Expression Omnibus (GEO) [17]. These data help to extract unique features of specific cell types, laying the foundation for estimating their fractions from bulk data.

Several cell type deconvolution methods have been proposed recently. Many researchers modelled this problem as a mixing process of biological signals from each cell type and employed overdetermined linear equations to solve it [18-22]. For example, CIBERSORT held bulk gene expression data as a linear mixture of different cell types [18]. The stability of the signature matrix was measured through the 2-norm condition number and support vector regression (SVR) was adopted to resolve the proportion of each cell type successfully [18]. Moss *et al*. proposed a linear model based on external references for DNA methylation data which successfully estimated the contribution of tissues to cfDNA [20]. Reference-free deconvolution methods such as non-negative matrix factorization (NMF) also perform well in some situations [23, 24]. Besides, some other studies modelled the mixing process with probabilistic models. For instance, PERT regarded the mixture of gene expression as a result of a sampling process and computed cell type proportions by the non-negative maximum likelihood model [25]. CancerLocator and CancerDetector regarded DNA methylation as a beta distribution to predict the fraction of tumor-derived reads in cfDNA sequencing data [26, 27].

It is rather remarkable that the above-mentioned studies did not pay sufficient consideration to rare cell types. Rare cell

types have relatively low proportions, therefore, deconvolution for these cell types are more prone to be biased by collinearity. In addition, these studies did not provide efficient methods to avoid the impact of outliers during the selection of cell-type specific markers, preventing themselves from accurate deconvolution [19, 28], especially for rare cell types.

Here, we presented ARIC as a new approach for robust and accurate inference of rare cell type proportions from bulk gene expression or DNA methylation data. ARIC adopts a novel two-step feature selection strategy to ensure an accurate and robust detection for rare cell types. ARIC introduces the component-wise condition number into eliminating collinearity step to pay equal attentions for the relative errors of all components. Besides, ARIC contains an automatic step for adaptively removing the outliers in markers, ensuring the robustness of the algorithm against noises. Finally, ARIC employs a weighted υ-support vector regression (υ-SVR) to get component proportions. We evaluated ARIC in DNA methylation and gene expression data from various experiments. ARIC outperforms other methods through many evaluation metrics regardless of data types, especially for the estimation of rare component proportions. These results demonstrate ARIC as an effective tool to infer cell type fractions, especially in scenarios where rare components are of vital importance.

Methods and Materials

Problem definition

First, we modelled the deconvolution problem as a linear mixture of different cell types as previous studies did [20, 22]. Here we used $\boldsymbol{m} \in \mathbb{R}^{N\times 1}$ and $\hat{\boldsymbol{m}} \in \mathbb{R}^{N\times 1}$ to denote the ideal value which do not contain any noise and the observed value of the bulk data, where N represents the number of markers. In order to estimate the proportion of cell types, we collected external references of possible components denoted as $\hat{\boldsymbol{X}}$, which is an $N \times K$ matrix where K is the number of cell types. The measured level of the i-th marker from the j-th cell type is denoted as $\hat{x}_{i,j}$ in the reference matrix $\hat{\boldsymbol{X}}$. Similarly, the ideal external reference is denoted as $\boldsymbol{X}$. The ideal proportion of different cell types in one sample is denoted as a non-negative vector $\boldsymbol{p} \in \mathbb{R}^{K\times 1}$, where the sum of $\boldsymbol{p}$ should be equal to 1. We use $\hat{\boldsymbol{p}}$ to denote the predicted value of $\boldsymbol{p}$. Under the assumption of linear mixing model, $\boldsymbol{m}$ can be regarded as the linear mixture of $\boldsymbol{X}$:

$$\boldsymbol{m} = \boldsymbol{X}\boldsymbol{p} \tag{1}$$

The task here is to estimate $\hat{\boldsymbol{p}}$ using $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{m}}$.

Employing component-wise condition number in eliminating feature collinearity

Solving $\boldsymbol{p}$ by Eq. 1 is an overdetermined equation problem due to the high-dimensional features of high-throughput omics data. Therefore, we first applied preliminary marker selection approaches to weaken the influence of biological and technical noise (see Supplementary Section 2). Even though, collinearity may occur in many features among different cell types, which may confuse the contributions of similar cell types and then lead to inaccurate proportion estimation. Previous studies [18, 29] used a condition-number-based strategy in RNA-seq data to measure the stability of the linear system against input variation or noise while reduce collinearity. As the measured $\boldsymbol{m}$ from a bulk sample is influenced by technical and biological noise, $\delta\boldsymbol{p}$ is defined as the change of proportion vector $\boldsymbol{p}$ if there occurs small changes or perturbations in the elements of $\boldsymbol{m}$, denoted as $\delta\boldsymbol{m}$. Besides, the relative errors of $\boldsymbol{p}$ and $\boldsymbol{m}$ are denoted as Δp and Δm respectively. Then the condition number can be denoted as C:

$$C := \max_{\delta\boldsymbol{m},\boldsymbol{m} \in \mathbb{R}^{N\times 1}} \frac{\Delta p}{\Delta m} \tag{2}$$

where

$$\Delta p = \frac{\|\delta\boldsymbol{p}\|}{\|\boldsymbol{p}\|}, \Delta m = \frac{\|\delta\boldsymbol{m}\|}{\|\boldsymbol{m}\|} \tag{3}$$

The condition number C measures the potential sensitivity of the proportion vector $\boldsymbol{p}$ to the change of bulk data $\boldsymbol{m}$ [30]. In particular, C takes the total relative error of the vector $\boldsymbol{p}$ into consideration, ignoring possible large relative errors on small-proportion components (see Supplementary Section 3).

In order to suppress the relative error of each component, here we used the component-wise condition numbers [31] to improve the procedure of marker selection. The definition of the component-wise condition number of the c-th component is:

$$CW_c := \max_{\delta\boldsymbol{m} \in \mathbb{R}^{N\times 1}} \frac{\Delta p_c}{\Delta m}, c \in \{1,2,\dots,K\} \tag{4}$$

where

$$\Delta p_c = \frac{\|\delta p_c\|}{\|p_c\|}, \Delta m = \frac{\|\delta\boldsymbol{m}\|}{\|\boldsymbol{m}\|} \tag{5}$$

Different from Eq. 2, the numerator in Eq. 4 changes from Δp to Δp_c, where Δp_c denotes the relative error of component c in the proportion vector $\boldsymbol{p}$.

However, directly computing CW_c in Eq. 4 is impractical since it is a theoretical deduction with unknown content Δp_c. Therefore, we used Eq. 6 below to calculate CW_c:

$$CW_c = \frac{\|\boldsymbol{t}_c\| \cdot \|\boldsymbol{m}\|}{|\boldsymbol{t}_c^T \cdot \boldsymbol{m}|} \tag{6}$$

where $\boldsymbol{t}_c^T \in \mathbb{R}^N$ refers to the c-th row of $\boldsymbol{X}^\dagger$ which denotes the pseudo-inverse of $\boldsymbol{X}$ [31]. We computing CW_c using $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{m}}$ in practice.

We further calculated the largest condition number among all K components as the upper bound of relative errors, denoted as CW_u:

$$CW_u = \max_{c \in \{1,2,\dots,K\}} CW_c \tag{7}$$

Since every component has a component-wise condition number, CW_u takes the most sensitive component into account. Through CW_u, we were able to select markers avoiding large relative errors in any component. One example in Supplementary Section 4 illustrates how component-wise condition numbers successfully represent the large relative error on small components.

We eliminated markers leading to strong collinearity and used component-wise condition numbers to avoid large relative errors on each component. This process includes three main steps: (i) inspired by a former study [20], we used Euclidean distances to find cell type pair with the strongest collinearity. (ii) Next, the most similar marker of the cell type pair was eliminated and CW_u was computed. We used Eq. 8 to find the most similar marker i_M in cell type pair j and k:

$$i_M = \arg\min_i \frac{|\hat{x}_{i,j} - \hat{x}_{i,k}|}{|\hat{x}_{i,k} + \hat{x}_{i,k}|} \tag{8}$$

(iii) The above two steps were repeated to find the markers corresponding to the minimum CW_u. It should be noted that searching the global minimum CW_u is impractical, therefore we adopted a heuristic procedure to tackle this problem (see Supplementary Section 8).

Adaptive and robust outlier detection

Through the above-mentioned marker selection procedure, we can select well-conditioned, non-collinear markers for each cell type. However, outliers that deviate from other markers may still exist, which may bring negative effects for precise deconvolution. Previously, few methods have tried to leverage outlier detection to improve the deconvolution performance. Here, we proposed a novel iterative and adaptive outlier detection method to overcome this problem based on robust regression [6, 32, 33].

We calculated the standardized residuals to indicate marker outliers. The standardized residual on the *i*-th marker is denoted as SR_i, which can be written as:

$$SR_i = \frac{r_i}{s} \tag{9}$$

where r_i is the prediction error for bulk data, denoted as:

$$r_i = \hat{m}_i - \sum_j \hat{x}_{i,j}\hat{p}_j$$

where $\hat{p}_j$ is the prediction value of the *j*-th component's fraction. In addition, s is defined as:

$$s = \sqrt{\frac{1}{N-K}\sum_{i=1}^{N} r_i^2}$$

Then a decision whether marker i should be regarded as an outlier is made:

$$D(i) = \begin{cases} 0, & |SR_i| \geq T \\ 1, & |SR_i| < T \end{cases} \tag{10}$$

The threshold T was used to decide which marker should be treated as outliers and was a fixed parameter determined at the start of the outlier-removal approach (see Supplementary Section 8). Next, markers detected as outliers were removed and standardized residuals were recalculated based on remaining markers. This approach was iterated for ten times with the fixed T, and the rest markers were used for proportion estimation.

Novel weighted SVR

Currently, υ-SVR has been proved to be a robust estimation method in many works [18, 34]. The primary problem of υ-SVR with linear kernel is as follows [35]:

$$\min_{\hat{\boldsymbol{p}}} \frac{1}{2}\hat{\boldsymbol{p}}^T\hat{\boldsymbol{p}} + C\left(\upsilon\epsilon + \frac{1}{N}\sum_{i=1}^{N}(\xi_i + \xi_i^*)\right) \tag{11}$$

$$\hat{x}_{i,\cdot}\hat{\boldsymbol{p}} - m_i \leq \epsilon + \xi_i$$

$$m_i - \hat{x}_{i,\cdot}\hat{\boldsymbol{p}} \leq \epsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0, i = 1, \ldots, N, \epsilon \geq 0$$

Here, $0 \leq \upsilon \leq 1$, C is the regularization parameter. $\hat{x}_{i,\cdot}$ denotes the i-th row of $\hat{\boldsymbol{x}}$, which represents the derived reference for marker i. The ϵ-insensitive loss function means that the loss is only considered when $\hat{x}_{i,\cdot}\hat{\boldsymbol{p}}$ is beyond the range of $m_i \pm \epsilon$. Note that we use m_i instead of $\hat{m}_i$ to denote the bulk data because $\hat{m}_i$ can be substituted by m_i after the previous feature selection.

From the Eq. 11, we can find that the absolute error term $|\hat{x}_{i,\cdot}\hat{\boldsymbol{p}} - m_i|$ not only determines which markers should be the support vectors but also influence the loss function value. Inspired by dampened weighted least squares (DWLS) which adjust the weights of markers in the absolute errors loss function [36], we proposed a novel deconvolution method integrating marker weights with υ-SVR to minimize the relative errors on each component and avoid the ignorance of relative errors on rare cell types.

The absolute error term can be written as:

$$absErr(\boldsymbol{m}, \hat{\boldsymbol{x}}, \hat{\boldsymbol{p}}) = \sum_{i=1}^{N} |m_i - \hat{x}_{i,\cdot}\hat{\boldsymbol{p}}| \tag{12}$$

Same as DWLS, we define $\hat{p}_j = \frac{p_j}{p_1}\hat{p}_1, j = 1,2,3\ldots,K$. Then, rewrite Eq. 12:

$$absErr(\boldsymbol{m}, \hat{\boldsymbol{x}}, \hat{\boldsymbol{p}}) = \sum_{i=1}^{N} |m_i - \sum_{j=1}^{K} \hat{x}_{i,j}\hat{p}_j|$$

$$= \sum_{i=1}^{N} |m_i - \hat{x}_{1,\cdot}\hat{p}_1 - \sum_{j=2}^{K} \hat{x}_{i,j}\frac{p_j}{p_1}\hat{p}_1|$$

$$= \sum_{i=1}^{N} |m_i - \sum_{j=1}^{K} \hat{x}_{i,j}p_j + \sum_{j=1}^{K} \hat{x}_{i,j}p_j - \hat{x}_{1,\cdot}\hat{p}_1 - \sum_{j=2}^{K} \hat{x}_{i,j}\frac{p_j}{p_1}\hat{p}_1|$$

$$= \sum_{i=1}^{N} |\sum_{j=1}^{K}(x_{i,j} - \hat{x}_{i,j})p_j + (p_1 - \hat{p}_1)(\hat{x}_{1,\cdot} + \sum_{j=2}^{K}\frac{\hat{x}_{i,j}p_j}{p_1})|$$

$$\geq \sum_{i=1}^{N} |(p_1 - \hat{p}_1)(\hat{x}_{1,\cdot} + \sum_{j=2}^{K}\frac{\hat{x}_{i,j}p_j}{p_1})| - \sum_{i=1}^{N} |\sum_{j=1}^{K}(x_{i,j} - \hat{x}_{i,j})p_j|$$

It is obvious that the second term is constant and the absolute error term is trying to minimize the errors on the first term. In consequence, we focus on the first term:

$$absErr(\boldsymbol{m}, \hat{\boldsymbol{x}}, \hat{\boldsymbol{p}}) \geq \sum_{i=1}^{N} |(p_1 - \hat{p}_1)(\hat{x}_{1,\cdot} + \sum_{j=2}^{K}\frac{\hat{x}_{i,j}p_j}{p_1})|$$

$$= \sum_{i=1}^{N} \left|\sum_{j=1}^{K} \hat{x}_{i,j}p_j\right| \left|\frac{p_1 - \hat{p}_1}{p_1}\right| \tag{13}$$

From Eq. 13, cell types which have larger reference value or a greater proportion lead to a larger impact on the absolute error term in υ-SVR and this phenomenon will lead to a biased estimation undoubtedly. Thus we designed weights for the markers to alleviate this problem:

$$w_i = \frac{1}{|\sum_{j=1}^{K} \hat{x}_{i,j}p_j|}$$

Then modify the absolute error term:

$$absErr(\boldsymbol{m}, \hat{\boldsymbol{x}}, \hat{\boldsymbol{p}}) = \sum_{i=1}^{N} w_i|m_i - \sum_{j=1}^{K} \hat{x}_{i,j}\hat{p}_j|$$

$$\geq \sum_{i=1}^{N} |\frac{p_1 - \hat{p}_1}{p_1}| = N|\frac{p_1 - \hat{p}_1}{p_1}|$$

Without the loss of generality, we have the following relationships

$$absErr(\boldsymbol{m}, \hat{\boldsymbol{x}}, \hat{\boldsymbol{p}}) = \sum_{i=1}^{N} w_i|m_i - \sum_{j=1}^{K} \hat{x}_{i,j}\hat{p}_j| \geq N|\frac{p_j - \hat{p}_j}{p_j}|$$

After weights adjustment, the absolute error term in υ-SVR can optimize the relative errors component-wisely, without ignoring rare cell types.

Finally, we normalize all the weights with Eq. 14 and apply υ-SVR to estimate proportions.

$$w_i = \frac{N * w_i}{\sum_{i=1}^{N} w_i}, w_i = \frac{1}{|\sum_{j=1}^{K} \hat{x}_{i,j} p_j|} \quad (14)$$

Evaluation Metrics

Here we used the root mean square error (RMSE) and the Pearson's correlation coefficient (PCC) as evaluation metrics as many former studies did [3, 18, 19]. Besides, in order to demonstrate the performance on different components, especially for rare components, we also adopted the mean absolute percentage error (MAPE) as another metric. The definition of RMSE and MAPE is:

$$\mathrm{RMSE}(\hat{\boldsymbol{p}}, \boldsymbol{p}) = \sqrt{\frac{1}{|K|}\sum_{i \in K}(\hat{p}_i - p_i)^2}$$

$$\mathrm{MAPE}(\hat{\boldsymbol{p}}, \boldsymbol{p}) = \frac{1}{|K|}\sum_{i \in K}\left|\frac{\hat{p}_i - p_i}{p_i}\right|$$

where $\hat{p}_i$ and p_i denotes the estimated value and the ground truth of each component respectively.

Benchmark methods

We selected state-of-art deconvolution methods that were designed for DNA methylation data or gene expression data for benchmark. Benchmarks for analyzing DNA methylation data include QP (quadratic programming) [37], Moss [20], Epidish (robust partial correlation) [38, 39], Sun [21], MethylCIBERSORT [40] and MethylResolver [41]. Benchmarks for analyzing gene expression data include QP [37], EPIC [19], CIBERSORT [18], dtangle [3], FARDEEP [6] and DeconRNASeq [42]. Of course, some methods which are already compared with other methods in previous works [19, 43] are not included in our analysis. Different preliminary marker selection strategies were used in different kinds of data (see Supplementary Section 2).

Datasets

We selected state-of-art HumanMethylationEPIC BeadArray data were used in both simulation and real data evaluation [44]. Data used for *in silico* simulation contains six cell types: NK cells, Neutrophils, B cells, monocytes, $CD4^+$ T cells and $CD8^+$ T cells, each cell type with six samples. Additionally, there are 12 real samples which are the mixtures of genomic DNA from six purified leukocyte subtypes, with known proportions for evaluation [44].

153 paired-end RNA-seq samples of 8 cell types were collected (Supplementary Table 2), including B cells, $CD4^+$ T cells, $CD8^+$ T cells, endothelial cells, macrophages, monocytes, neutrophils and NK cells. We processed raw RNA-seq data into transcripts per million (TPM) to generate simulation data [19]. Single-cell RNA-seq (scRNA-seq) data were adopted from a previous study [45]. We selected scRNA-seq data of 7 tissues from 6 to 10-week-old female mice for generating pseudo-bulk data.

Four gene expression microarray datasets and three RNA-seq datasets with known proportion are collected.

All the datasets can be accessed through supplementary table 1~4.

Results

ARIC infers the cell type fraction accurately on in silico mixed data

We first evaluated ARIC on different *in silico* mixed datasets including DNA methylation assays, RNA-seq and scRNA-seq. Each dataset was divided into two parts, one for constructing the external reference and the other for evaluation. Proportions of all components were generated randomly and adequate samples were produced to ensure the existence of rare components (see Supplementary Section 7). RMSE, MAPE and PCC were calculated to evaluate the performance of different methods (Fig. 1).

Though PCCs of some methods are comparable with ARIC, ARIC achieves a relatively low RMSE and MAPE compared with any other methods for both DNA methylation assays (Fig. 1A-B) and RNA-seq (Fig. 1E-F), which indicates the high accuracy of ARIC for the prediction of every component's fraction. What's more, as shown in Fig. 1C~D and G~H, the prediction results of ARIC are squeezed into the diagonal line gradually as the true fraction decreased by contrast with other methods, which indicates the precision of ARIC on rare components. Interestingly, the prediction results of dtangle showed a significant deviation in comparison with the ground truth (the last panel in Fig. 1G), though with a high PCC.

To further evaluate the performance of ARIC, we generated pseudo-bulk RNA-seq samples *in silico* with data adopted from a well-characterized scRNA-seq study [45]. As shown in Supplementary Fig. S1, ARIC outperforms in metrics compared with others, especially for rare components. Interestingly, the deviation of dtangle's prediction still exists. dtangle shows a higher PCC as well as a smaller RMSE than all the other methods except for ARIC and FARDEEP, but exhibits the highest MAPE value in the meantime. The results, together with the results shown in Fig. 1G, indicate that PCC and RMSE are insufficient to depict the prediction accuracy and may cause erroneous judgements.

ARIC estimate rare cell type fractions accurately

To illustrate the capability of ARIC to deal with rare components, we computationally mixed methylation data [20] of individual cell types at varying proportions. *In silico* simulations were performed by setting the proportion of one component to 1%, 3%, 5%, 7% and 10% respectively, and the proportions of other components were generated randomly. We evaluated the results of the rare component solely with all the three metrics (Fig. 2). ARIC always shows the smallest RMSE (Fig. 2A) and MAPE (Fig. 2B) among all methods across different rare-component proportions. Moreover, the results of ARIC are the closest to the ground truth with the smallest variance across different proportions (Fig. 2C), which illustrates the robustness of ARIC on inferring the fraction of rare components. The same results were obtained when inferring components with lower proportions from 0.1% to 1% (Supplementary Fig. S2).

ARIC outperforms in the deconvolution of real data from multiple sources

To evaluate the efficacy of ARIC on real data, we collected a DNA methylation dataset [44], which provides us 12 samples with known fractions of each cell type. The deconvolution results are shown in Fig. 3. Though all methods exhibit high

PCCs, ARIC shows a better performance than any other baseline methods on the metrics of both RMSE and MAPE.

We also collected seven expression datasets measured by microarray or RNA-seq with known cell type fractions. We calculated the performance of each method on all datasets and analyzed the results jointly. ARIC achieves a relatively lower RMSE (Fig. 4A) and MAPE (Fig. 4B) in different datasets. The median RMSE and MAPE of dtangle are almost the same as ARIC, but our method shows more outstanding outcomes when focusing on rare components (Fig.4C-D). ARIC also shows the smallest deviation among all methods (Fig. 4), which indicates the robustness of ARIC.

Acknowledgements

We would like to thank Dr. Qiongye Dong and Ms. Jiaqi Li for helpful discussions.

Discussion

As a new approach for deconvolution of cell type fractions, ARIC is tested on both *in silico* mixed data as well as real data with various DNA methylation and gene expression datasets. Several benchmarks are selected to compare with ARIC. For all kinds of datasets, ARIC shows the ability to estimate the fractions of different cell types, especially for rare components, more precisely than any other deconvolution methods.

The remarkable performance of ARIC owes to several novel designs in the algorithm. Different cell types that are differentiated from same progenitors or share similar functions may exhibit similar methylome or transcriptome profiles, which may further lead to confounded deconvolution results due to collinearity [36]. Previous studies minimized the condition number to improve the accuracy of the deconvolution results [18, 34]. However, the condition number cannot pay equal attention to all cell types as it measures the overall error of all components instead of the component-wise error. This will cause a strong bias for rare cell types. Therefore, the component-wise condition number is introduced into ARIC to evaluate the error of each component more equally to get rid of the negligence of rare components. During the marker selection process, the component-wise condition number is calculated at each step, and then markers leading to the smallest component-wise condition number are adopted for further analysis. The deconvolution results of both simulated and real datasets reveal that employing the component-wise condition number brings ARIC a more powerful capacity to estimate the fraction of rare components.

Furthermore, robustness is an indispensable requirement for bulk data deconvolution to ensure high accuracies. However, the deconvolution procedure is susceptible to outliers brought by measure error or environmental effect, which may hamper the robustness of algorithms [6]. Here, ARIC utilizes the standardized residual to distinguish outliers from effective markers. Notably, some outliers are hard to be differentiated when there exist more significant outliers in all markers. As the consequence, outliers can hardly be detected and removed without iterations. Therefore, ARIC computes standardized residuals and detects outliers adaptively to ensure that outliers are removed as precisely as possible.

There is still room for improvement of ARIC. ARIC depends on external references whose quality may influence the prediction results significantly. As the rapid development of single-cell sequencing technologies, purer references can be produced to enhance the performance of ARIC. Moreover, as ARIC is developed independent from the type of data and performs well in data generated from HumanMethylationEPIC BeadArray, microarray, RNA-seq and scRNA-seq, ARIC is promising to be applied into other bulk data, such as ATAC-seq and FAIRE-seq.

In conclusion, ARIC is a robust and accurate tool for decoupling the fraction of cell types in mixture data. Particularly, ARIC can estimate the fraction of rare components far more precisely and robustly than other methods, suggesting ARIC as a promising tool to solve the deconvolution problem of bulk data where rare components matter, which may further benefit both scientific research and clinical applications.

References

1. Feng, H. and H. Wu, Differential methylation analysis for bisulfite sequencing using DSS. Quantitative Biology, 2019. 7(4): p. 327-334.
2. Ismail, W.M., E. Nzabarushimana, and H. Tang, Algorithmic approaches to clonal reconstruction in heterogeneous cell populations. Quantitative Biology, 2019: p. 1-11.
3. Hunt, G.J., et al., dtangle: accurate and robust cell type deconvolution. Bioinformatics, 2019. 35(12): p. 2093-2099.
4. Shen-Orr, S.S., et al., Cell type–specific gene expression differences in complex tissues. Nature methods, 2010. 7(4): p. 287-289.
5. Galon, J., et al., Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science, 2006. 313(5795): p. 1960-1964.
6. Hao, Y., et al., Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. PLoS Comput Biol, 2019. 15(5): p. e1006976.
7. Mlecnik, B., et al., Integrative analyses of colorectal cancer show immunoscore is a stronger predictor of patient survival than microsatellite instability. Immunity, 2016. 44(3): p. 698-711.
8. Fridman, W.H., et al., The immune contexture in human tumours: impact on clinical outcome. Nature Reviews Cancer, 2012. 12(4): p. 298-306.
9. Gentles, A.J., et al., The prognostic landscape of genes and infiltrating immune cells across human cancers. Nature medicine, 2015. 21(8): p. 938-945.
10. Lo, Y.D., et al., Presence of fetal DNA in maternal plasma and serum. The lancet, 1997. 350(9076): p. 485-487.
11. Schwarzenbach, H., D.S. Hoon, and K. Pantel, Cell-free nucleic acids as biomarkers in cancer patients. Nature Reviews Cancer, 2011. 11(6): p. 426-437.
12. Fiala, C. and E.P. Diamandis, Utility of circulating tumor DNA in cancer diagnostics with emphasis on early detection. BMC medicine, 2018. 16(1): p. 166.
13. Robins, H.S., et al., Digital genomic quantification of tumor-infiltrating lymphocytes. Science translational medicine, 2013. 5(214): p. 214ra169-214ra169.
14. Saltz, J., et al., Cancer Genome Atlas Research N, Shmulevich I. AUK R, Lazar AJ, Sharma A, Thorsson, 2018. 2018: p. 181-193.
15. Network, C.G.A., Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature, 2015. 517(7536): p. 576-582.
16. Consortium, E.P., An integrated encyclopedia of DNA elements in the human genome. Nature, 2012. 489(7414): p. 57-74.
17. Edgar, R., M. Domrachev, and A.E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 2002. 30(1): p. 207-210.
18. Newman, A.M., et al., Robust enumeration of cell subsets from tissue expression profiles. Nat Methods, 2015. 12(5): p. 453-7.
19. Racle, J., et al., Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife, 2017. 6.
20. Moss, J., et al., Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun, 2018. 9(1): p. 5068.

21. Sun, K., et al., Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proc Natl Acad Sci U S A, 2015. 112(40): p. E5503-12.
22. Feng, H., P. Jin, and H. Wu, Disease prediction by cell-free DNA methylation. Brief Bioinform, 2019. 20(2): p. 585-597.
23. Tang, D., S. Park, and H. Zhao, NITUMID: Nonnegative matrix factorization-based Immune-TUmor MIcroenvironment Deconvolution. Bioinformatics, 2020. 36(5): p. 1344-1350.
24. Houseman, E.A., et al., Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC bioinformatics, 2016. 17(1): p. 259.
25. Qiao, W., et al., PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. PLoS Comput Biol, 2012. 8(12): p. e1002838.
26. Li, W., et al., CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. Nucleic Acids Res, 2018. 46(15): p. e89.
27. Kang, S., et al., CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. Genome Biol, 2017. 18(1): p. 53.
28. Miao, Y.R., et al., ImmuCellAI: A Unique Method for Comprehensive T‐Cell Subsets Abundance Prediction and its Application in Cancer Immunotherapy. Advanced Science, 2020. 7(7).
29. Chen, S.-H., et al., A gene profiling deconvolution approach to estimating immune cell composition from complex tissues. BMC bioinformatics, 2018. 19(4): p. 154.
30. Belsley, D.A., E. Kuh, and R.E. Welsch, Regression diagnostics: Identifying influential data and sources of collinearity. Vol. 571. 2005: John Wiley & Sons.
31. Winkler, J., A statistical analysis of the numerical condition of multiple roots of polynomials. Computers & Mathematics with Applications, 2003. 45(1-3): p. 9-24.
32. Huang, D., R. Cabral, and F. De la Torre, Robust regression. IEEE transactions on pattern analysis and machine intelligence, 2015. 38(2): p. 363-375.
33. Rousseeuw, P.J. and A.M. Leroy, Robust regression and outlier detection. Vol. 589. 2005: John wiley & sons.
34. Li, H., et al., DeconPeaker, a deconvolution model to identify cell types based on chromatin accessibility in ATAC-Seq data of mixture samples. Frontiers in genetics, 2020. 11: p. 392.
35. Chang, C.-C. and C.-J. Lin, Training v-support vector regression: theory and algorithms. Neural computation, 2002. 14(8): p. 1959-1977.
36. Tsoucas, D., et al., Accurate estimation of cell-type composition from gene expression data. Nat Commun, 2019. 10(1): p. 2975.
37. Gass, S.I. and C.M. Harris, QP, in Encyclopedia of Operations Research and Management Science, S.I. Gass and C.M. Harris, Editors. 2001, Springer US: New York, NY. p. 655-655.
38. Zheng, S.C., et al., EpiDISH web server: Epigenetic Dissection of Intra-Sample-Heterogeneity with online GUI. 2020, Oxford University Press.
39. Teschendorff, A.E., et al., A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC bioinformatics, 2017. 18(1): p. 1-14.
40. Chakravarthy, A., et al., Pan-cancer deconvolution of tumour composition using DNA methylation. Nature communications, 2018. 9(1): p. 1-13.
41. Arneson, D., X. Yang, and K. Wang, MethylResolver—a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents. Communications biology, 2020. 3(1): p. 1-13.
42. Gong, T. and J.D. Szustakowski, DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics, 2013. 29(8): p. 1083-5.
43. Sturm, G., et al., Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics, 2019. 35(14): p. i436-i445.
44. Salas, L.A., et al., An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol, 2018. 19(1): p. 64.
45. Han, X., et al., Mapping the Mouse Cell Atlas by Microwell-Seq. Cell, 2018. 172(5): p. 1091-1107 e17.

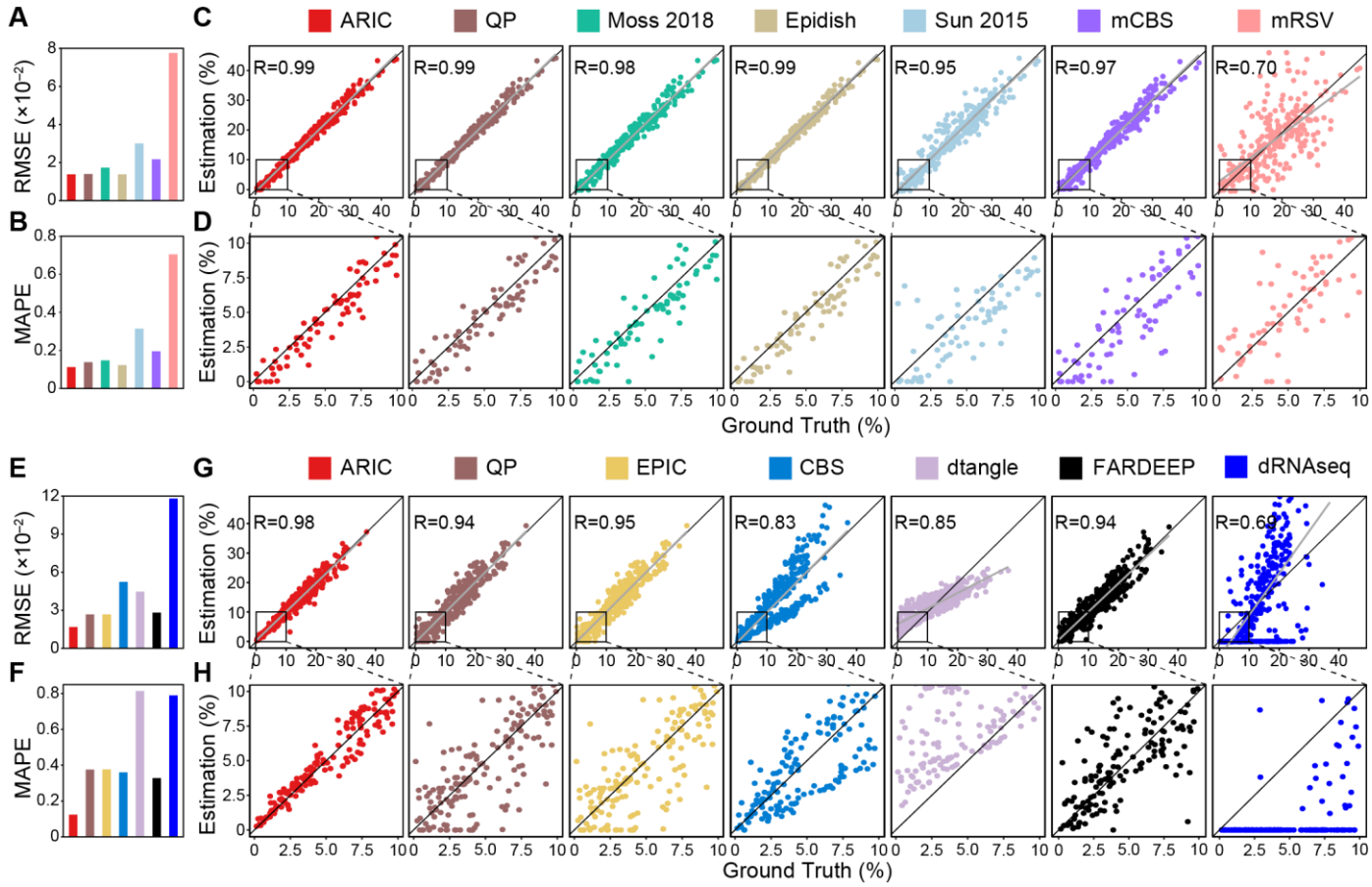

Fig. 1 Deconvolution results for *in silico* bulk DNA methylation (A~D) and gene expression (E~H) datasets. (A) and (E) the RMSE for each method. (B) and (F) the MAPE for each method. (C) and (G) are the scatter plots for estimated cell-type fractions against true fractions. PCCs are shown in the top-left corner of each panel. (D) and (H) are zoomed-in versions of (C) and (G) for cell-types with fraction less than 10%. Gray lines represent linear regressions of the points. mCBS: MethylCIBERSORT, mRSV: MethylResolver, CBS: CIBERSORT, dRNAseq: DeconRNASeq.

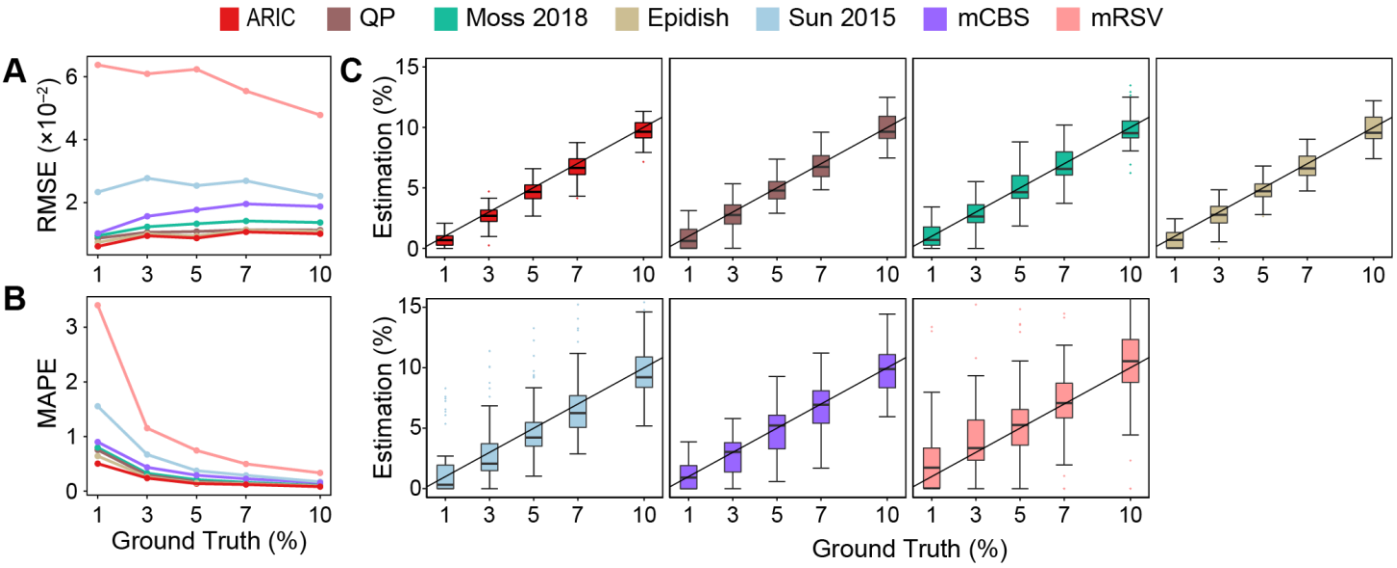

Fig. 2 Deconvolution results for simulated rare components varying from 1% to 10%. (A) and (B) show RMSEs and MAPEs gradually changing with the rare component proportion. (C) is the box-plot of the deconvolution results with replicates (n = 60). Black lines represent the estimation values are equal to the ground truth. mCBS: MethylCIBERSORT, mRSV: MethylResolver.

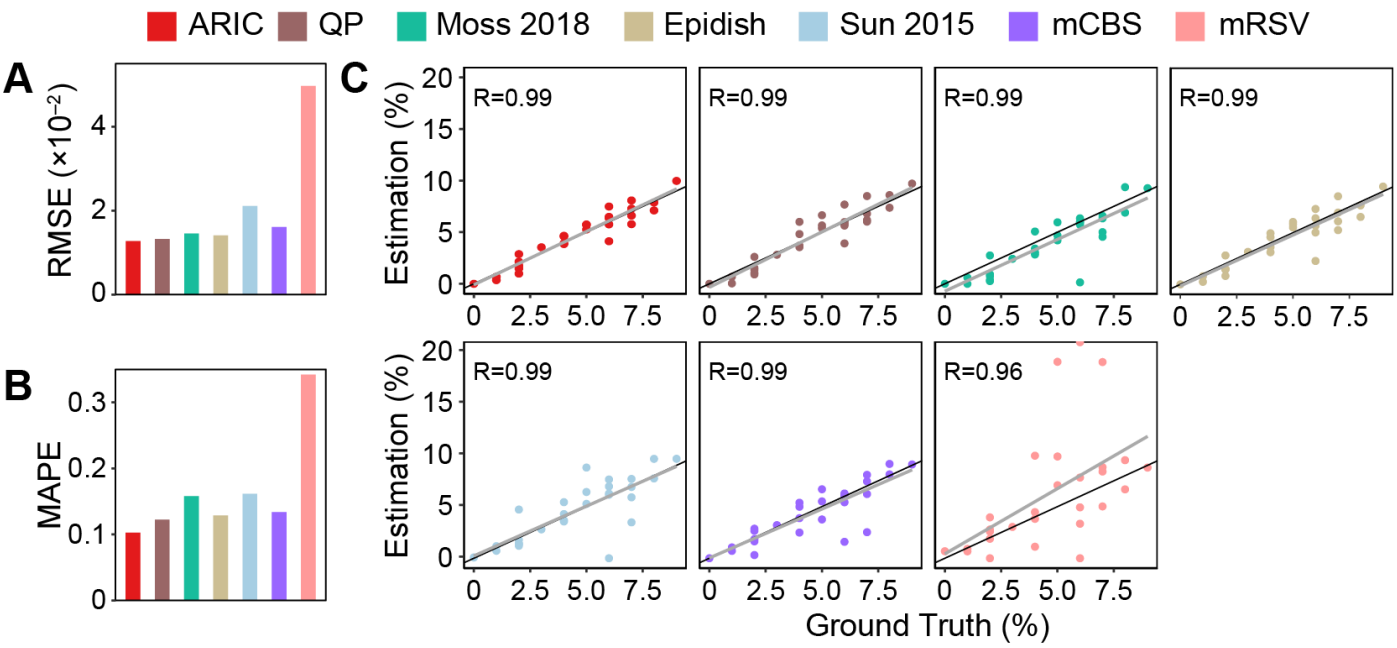

Fig. 3 Deconvolution results for experimentally mixed methylation data. (A~C) show RMSEs, MAPEs and prediction results against the ground truth respectively. PCCs are shown in the top-left corner for each method in (C). Gray lines in (C) represent linear regressions of the points. Black lines in (C) represent the estimation values are equal to the ground truth. mCBS: MethylCIBERSORT, mRSV: MethylResolver.

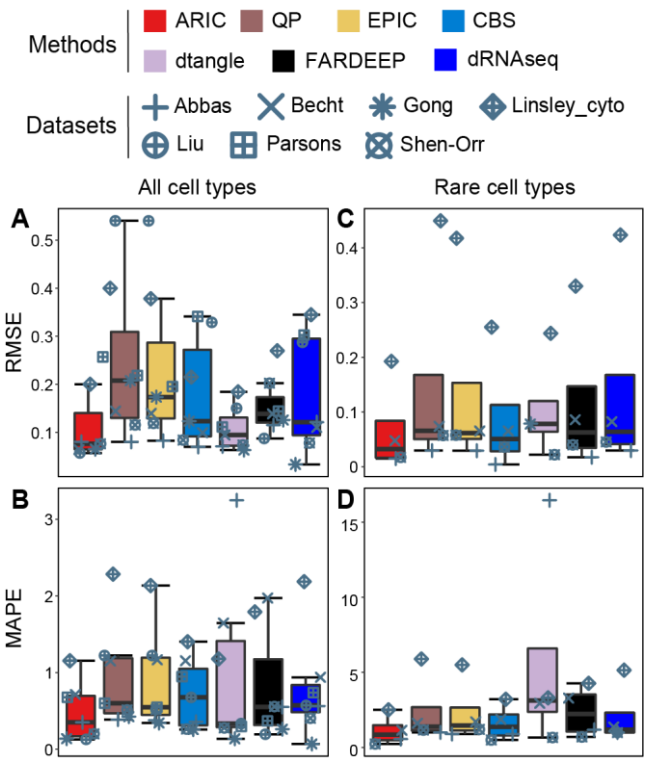

Fig. 4 Deconvolution results for experimentally mixed expression data. (A) and (B) are RMSEs and MAPEs for all components. (C) and (D) are RMSEs and MAPEs for rare components with fractions less than 10%. Black lines represent the median for each method. CBS: CIBERSORT. dRNAseq: DeconRNASeq.