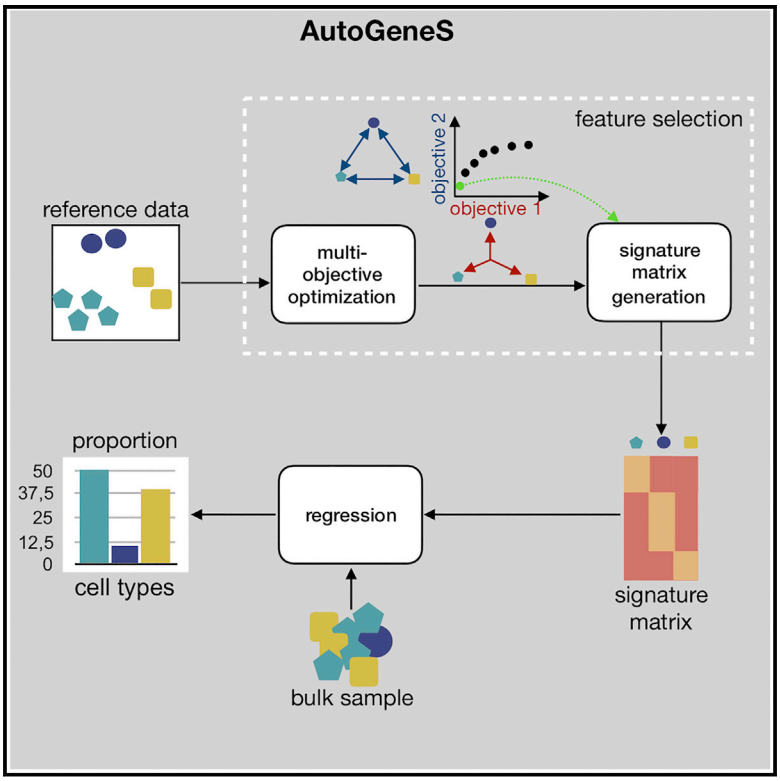
Cell Systems

Methods

# AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution

## Graphical Abstract




## Highlights

- Automatically learns informative genes for bulk RNA-seq deconvolution
- Proposes a multi-objective optimization approach for feature selection
- Uses reference profiles from single-cell data or sorted cell populations
- Is easily extendable to user-specific objectives or regression methods

## Authors

Hananeh Aliee, Fabian J. Theis

## Correspondence

fabian.theis@helmholtz-muenchen.de

## In brief

Aliee et al. proposed a novel feature selection method integrated into a deconvolution platform, called AutoGeneS, to infer cellular proportions from a bulk RNA-seq sample. This study enables the exploration of the tremendous quantity of biomedical data that has been collected in clinics and population cohorts but is yet under-explored.

Aliee & Theis, 2021, Cell Systems *12*, 706–715
July 21, 2021 © 2021 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2021.05.006

CellPress

CellPress
OPEN ACCESS

Cell Systems

Methods

# AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution

Hananeh Aliee[1] and Fabian J. Theis[1,2,3,4,*]
[1]Institute of Computational Biology, Helmholtz Centre, Munich, Bayern 85764, Germany
[2]Department of Mathematics, Technical University of Munich, Munich, Bayern 85748, Germany
[3]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany
[4]Lead contact
*Correspondence: fabian.theis@helmholtz-muenchen.de
https://doi.org/10.1016/j.cels.2021.05.006

## SUMMARY

Knowing cell-type proportions in a tissue is very important to identify which cells or cell types are targeted by a disease or perturbation. Hence, several deconvolution methods have been proposed to infer cell-type proportions from bulk RNA samples. Their performance with noisy reference profiles and closely correlated cell types highly depends on the set of genes undergoing deconvolution. In this work, we introduce AutoGeneS, a platform that automatically extracts discriminative genes and reveals the cellular heterogeneity of bulk RNA samples. AutoGeneS requires no prior knowledge about marker genes and selects genes by simultaneously optimizing multiple criteria: minimizing the correlation and maximizing the distance between cell types. AutoGeneS can be applied to reference profiles from various sources like single-cell experiments or sorted cell populations. Ground truth cell proportions analyzed by flow cytometry confirmed the accuracy of AutoGeneS in identifying cell-type proportions. AutoGeneS is available for use via a standalone Python package (https://github.com/theislab/AutoGeneS).

## INTRODUCTION

Bulk RNA samples are routinely collected and profiled for clinical purposes and biological research to study gene expression patterns in various conditions such as disease states. Such samples reflect mean gene expression across thousands of cells and, thus, mask cellular heterogeneity in a complex tissue. However, the knowledge of cell-type composition and their fractions helps to characterize molecular changes in diseased tissues that are important for the identification of disease-related cell types as well as the development of targeted drugs and therapies (Kuhn et al., 2011). Studying the variation of cell-type composition also opens new avenues in analyzing a tremendous yet underexplored quantity of biomedical data that has already been collected in clinics. Therefore, a number of deconvolution techniques have been proposed in the literature for analyzing cellular composition from mixture samples (Du et al., 2019), (Frishberg et al., 2019), (Hunt et al., 2019), (Kuhn et al., 2012), (Shen-Orr and Gaujoux, 2013), (Schwartz and Shackney, 2010), (Zaitsev et al., 2019). These techniques mainly rely on a so-called signature matrix, the mean expression of signature (or marker) genes chiefly of well-defined cell types (Baron et al., 2016), (Kuhn et al., 2012). Some of these techniques are only applicable to the samples of a specific tissue and use a predefined signature matrix for deconvolution (Aran et al., 2017), (Kuhn et al., 2011), (Monaco et al., 2019), (Schelker et al., 2017), (Shen-Orr et al., 2010). However, due to the advances in single-cell RNA-sequencing (scRNA-seq), many cell subtypes are still being discovered whose specific signatures are yet unknown. Therefore, automatic techniques are needed to specify cell-type-specific signatures, i.e., determination of a subset of transcripts also known as feature selection, in an unsupervised manner without requiring prior knowledge about the markers (Figure 1A). The signatures can then be employed to generate a signature matrix from the reference dataset. To fill this need, previous studies have explored general feature selection techniques that can be extended to almost any tissue type (Du et al., 2019), (Kang et al., 2018), (Newman et al., 2019), (Wang et al., 2019). These techniques typically rank genes in each population (or cell type) based on a single-criterion comparison (often q value from a t test) and select the top-ranked genes that differentiate one population from the others. However, when complex tissues contain highly similar cell types and, more specifically, when the reference profiles are noisy (as is often the case for scRNA-seq data; Brennecke et al., 2013), the signatures must be selected based on additional criteria. Moreover, these techniques often perform feature selection on each population individually and accumulate population-specific features. This can result in a large, overlapping set of genes between closely related

 © 2021 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

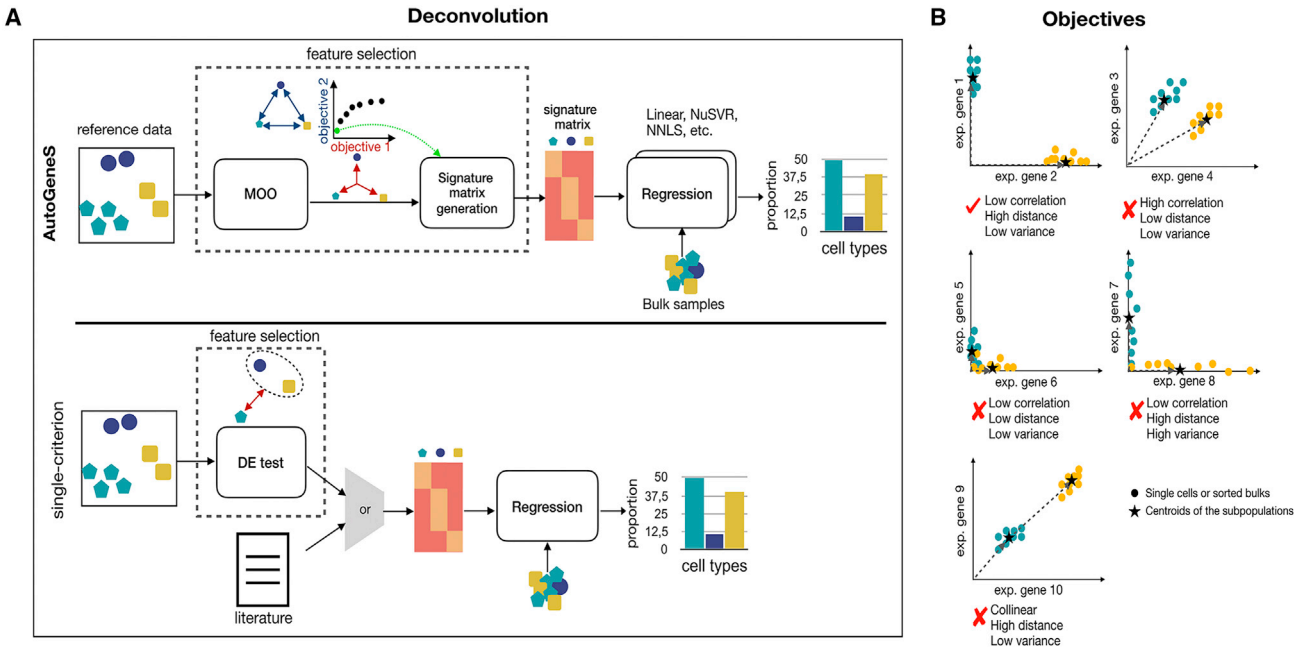



**Figure 1. Framework for dissecting cellular composition**

(A) AutoGeneS starts from the reference data of the populations (cell types) and (I) finds the discriminative features using a multi-objective optimization (MOO) method, (II) generates a signature matrix by subsetting the reference profiles, and (III) calls one of the built-in regression methods to infer the cellular proportions. Conventional methods, either use tissue-specific signatures driven from literature or perform a single-criterion test, like a t test (comparing one population with the rest), to find differentially expressed genes.

(B) We indicate through five examples that minimizing inter-population correlation is not necessarily sufficient for feature selection. Gene 1 and gene 2 are expressed in only one of the two populations with a high distance and a low intra-population variance. Gene 3 and gene 4 are correlated, the centroids of the populations are very close for gene 5 and gene 6, both gene 7 and gene 8 have high variances, and finally, the populations are collinear (also highly correlated) for the set of gene 9 and gene 10.

populations. Therefore, it is important to globally select a set of signatures that discriminates all populations.

In this study, we present a platform, named AutoGeneS that automatically selects differentiating signatures for the purpose of bulk deconvolution. To accomplish this, we developed a feature selection method employing multi-objective optimization that leverages reference profiles from single-cell or bulk-sorted transcriptional data. The optimization approach targets closely related cell types and aims at selecting a set of genes that simultaneously minimizes the correlation and maximizes the distance between the cell types. The proposed feature selection is robust against technical and biological noise and, together with a regression method, can reveal cell proportion of complex mixture samples without any a priori knowledge about the marker genes. Several regression methods are integrated into AutoGeneS that use the automatically generated signature matrix to infer cellular proportions. Through comprehensive benchmark evaluations and applications to peripheral blood mononuclear cells (PBMC) in humans with single-cell and bulk-sorted reference profiles as well as the ethmoid sinus of patients with and without nasal polyposis, we show that AutoGeneS outperforms other approaches. Moreover, when the reference data for feature selection includes more than one sample, we show that the genes selected by AutoGeneS reduce batch effects. This is of utmost importance when reference datasets include many samples that are captured at different times and/or generated in multiple labs using different technology platforms (Tran et al., 2020).

## RESULTS

### AutoGeneS enables tissue characterization using reference profiles across multiple technologies

To dissect cellular content from bulk RNA samples, we first require reference gene expression data, which may originate from various technologies such as single-cell experiments or sorted cell populations. For scRNA-seq data, dimensionality reduction and clustering are first performed to reveal cell types and annotate cells. AutoGeneS takes the annotated reference data and generates the reference profiles as the closest points to the centroids of the cell types (see STAR methods for technical details). Depending on the source of data, we refer to them as either single-cell or sorted bulk reference profiles. AutoGeneS then identifies discriminative genes for deconvolution using multi-objective optimization (MOO) (Figure 1A). This is particularly crucial for (1) reducing computational complexity by eliminating unexpressed genes or genes with a uniform expression level across the cell types that are important for large-scale studies and (2) increasing the signal-to-noise ratio by selecting genes that differentiate the cell types. Once we filtered the reference profiles for differentially expressed genes, called a

signature matrix, we leveraged them to predict cell proportions from other mixture profiles.

Previous studies show that the expression profiles from each cell type are linearly additive; this makes the contribution of each cell type proportional to its fraction in the mixture profile (Zaitsev et al., 2019). While several regression methods are integrated into AutoGeneS, we employed Nu-Support Vector Regression (Nu-SVR) in this paper to deconvolve a given mixture as the last step of the workflow. This makes the regression robust against technical and biological noise by discarding a certain fraction of outliers specified by Nu and performing regression on the remaining samples (Chen et al., 2005), (Chang and Lin, 2011), (Newman et al., 2015) (discussed in STAR methods).

## Multi-objective optimization can learn non-collinear genes

In this study, we propose a MOO approach to select a set of non-collinear genes and show that excluding collinear genes significantly improves the accuracy of the inferred proportions. Collinearity occurs when two or more predictor variables (cell-type-specific profiles) in a statistical model are linearly correlated. In the case of high collinearity, coefficient estimates (the proportions of cell types) are highly sensitive to small changes in the dataset. Because biological datasets (specifically scRNA-seq data) are typically considered noisy, collinearity is one of the main challenges in bulk deconvolution. Several quantifications of collinearity are proposed in the literature, with the most common being the pair-wise correlation coefficient (Dormann et al., 2013). While correlation and collinearity are not equivalent, high absolute correlation coefficients are usually indicative of high linear relatedness. AutoGeneS is then developed to select a set of genes that minimizes the global correlation between all the predictors simultaneously to reduce collinearity. By global correlation, we mean the sum of the pair-wise correlation coefficients between all pairs of populations.

However, selecting a subset of genes that only minimizes the correlation might not be enough (Figure 1B). Indeed, when the Euclidean distance between the centroids of two populations is low, the coefficient estimates become highly sensitive to noise in the dataset (Figure 1B, genes $g_5$ and $g_6$). Moreover, selecting unstable genes with high inter-cluster variance can result in low representability of the centroids, which can increase regression error (Figure 1B, genes $g_7$ and $g_8$). Therefore, rather than selecting genes based only on a pair-wise correlation coefficient, we propose a MOO approach that simultaneously minimizes global correlation and maximizes the sum of pair-wise Euclidean distances (Figures 1A and S1). We may also consider the inter-cluster variance as a third objective to be minimized, but instead, we recommend adding this as a constraint to the optimizer by filtering out genes with high inter-cluster variance (for more discussion, see STAR methods).

For a MOO problem, there usually exists no single solution that simultaneously optimizes all objectives. In this case, the objective functions are said to be conflicting, and there exists a (possibly infinite) number of Pareto-optimal solutions (Figure 1A). Pareto-optimal solutions are a set of all solutions that are not dominated by any other explored solution. Optimization-wise, Pareto-optimal solutions offer a set of equally good solutions from which to select, depending on the dataset (discussed in STAR methods). After selecting a solution, the reference profiles are filtered for the corresponding genes to generate the signature matrix (Figure 1A). To visualize Pareto-optimal solutions, we applied AutoGeneS to the human embryonic scRNA-seq data from (Chu et al., 2016) using 4,000 preselected highly variable (HV) genes (Figure S2). Conventional techniques typically use a t test between each population and all other populations to obtain differentially expressed genes (Figure 1A). We compared the objective values of the solutions from AutoGeneS with differentially expressed genes using t test (T-DE), HV genes, and a set of random solutions, for which some genes are randomly selected as markers. The results show that solutions from AutoGeneS dominate other solutions and, moreover, AutoGeneS is capable of finding solutions with a very low global correlation and a relatively high global distance (Figure S2B).

## AutoGeneS selects non-batch driver genes

When reference datasets include multiple samples, it is important to ensure that selected genes for deconvolution are not exposed to large variations or batch effects. Otherwise, batch effects vary the signature matrix and the deconvolution results, respectively. To address this concern, we performed different feature selection methods on two datasets with multiple samples: healthy pancreas (n = 6 samples and 3 studies) and fetal pancreas (n = 2 samples) (Baron et al., 2016), (Enge et al., 2017), (Segerstolpe et al., 2016), (Han et al., 2020). Healthy pancreas data with a total of 10,955 cells includes one inDrop and two Smart-seq2 datasets. Fetal pancreas data consists of two microwell-seq datasets with a total of 9,456 cells. We treated each sample as a batch and used a scaled Silhouette score to evaluate the batch effects for the set of genes selected by each feature selection method. Here, we compared AutoGeneS (n = 400 genes) to HV genes (n = 4,000), T-DE genes (n = 400), and CIBERSORTx (n = 3,203 genes for healthy and n = 2,109 genes for fetal pancreas) (Newman et al., 2019) (Figure S3). Scaled Silhouette score of 1 represents ideally mixed batches and strongly separated cell types (see STAR methods) (Büttner et al., 2019). Results show that in both datasets, AutoGeneS achieved the highest scaled Silhouette score and the selected genes reduced batch effects (Figure S3). While T-DE genes perform similarly to AutoGeneS for healthy pancreas, we show later that they are not sufficiently good for deconvolution because they do not necessarily reduce cellular collinearity.

## Evaluation on simulated bulk tissues

First, to systematically benchmark, we applied AutoGeneS to synthetic bulk RNA-seq data. The reference data for signature learning was human embryonic scRNA-seq data from Chu et al. (2016). To make the deconvolution challenging but also more realistic, the synthetic bulk RNA samples were generated by summing the matched sorted bulk RNA-seq read counts of the same tissue from (Chu et al., 2016). Indeed, library preparation protocols vary across different sequencing technologies as well as laboratories, and bulk deconvolution analysis therefore may be performed on reference data generated differently in either way. In this scenario, true cell-type proportions are known, which allows a comparison of the accuracy of the

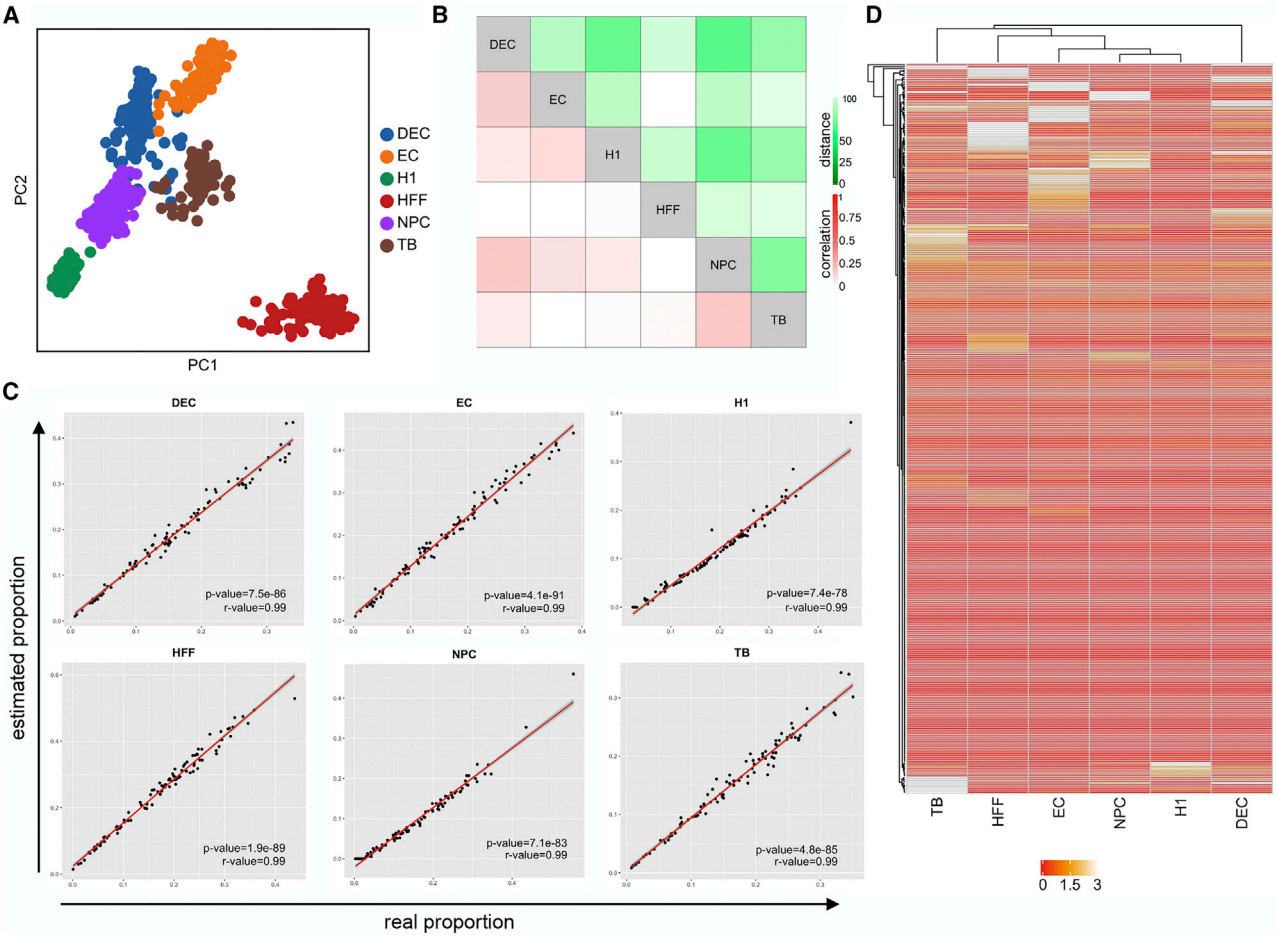



**Figure 2. Bulk deconvolution of synthetic bulk RNA samples with human embryonic single-cell reference profiles**

The bulk data is constructed using sorted bulk RNA-seq read counts while the matched single-cell reference data is taken from Chu et al. (2016).

(A) Principal component analysis (PCA) projection of the scRNA-seq data with six major populations. Shown is PC1 versus PC2.

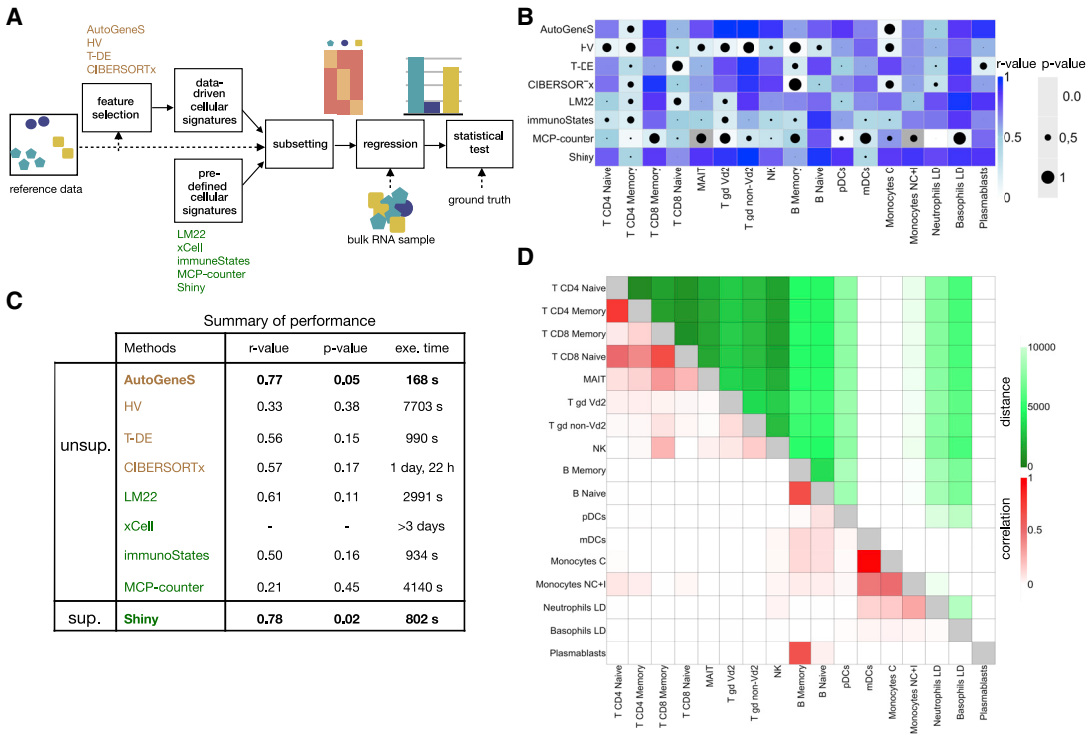
(B) Shown are pair-wise correlation and distance between the cell types using the signature matrix inferred from the optimization.

(C) Scatterplots illustrating real and estimated proportions for 100 synthetic bulk RNA samples. The significance of the results was assessed by r values and p values from linear regression (solid red lines).

(D) Heatmap of signature matrix inferred by AutoGeneS. NPC, neuronal progenitor cell; DEC, definitive endoderm cell; EC, endothelial cell; TB, trophoblast-like cell; HFF, human foreskin fibroblasts.

proposed technique. In this experiment, AutoGeneS searched for 400 marker genes that minimized cosine similarity (as a measure of correlation) and maximized Euclidean distance between six major cell types (Figure 2). After filtering the 400 marker genes with the lowest mean correlation coefficients among all the Pareto-optimal solutions, the average pair-wise correlation coefficient was equal to 0.07. This shows that AutoGeneS efficiently explores the search space seeking non-collinear genes (Figure 2C). Moreover, compared with ground truth cell proportions, the deconvolution results show high accuracy, with a r value > 0.99 and a p value < 1e-78 (Figure 2C). The heatmap of the signature matrix further indicates that the optimization result converges to the genes that are either highly expressed in only one cell type or lowly but differentially expressed between the cell types (Figure 2D). See the STAR methods section for more details on how to select the number of marker genes using AutoGeneS.

To study the robustness of AutoGeneS to both cellular correlation and noise in gene expression, we removed top differentially expressed genes (n = 5,379) as well as HV genes (n = 5,000) and incrementally added to the noise level through randomly dropping out genes in the dataset. Removing differentially expressed and HV genes increased the average cellular correlation by 118% for HV genes (Figure S4A). However, AutoGeneS still returned very accurate predictions (r value = 0.89 and p value = 3.11E-20). In addition, by increasing the noise level up to 50%, AutoGeneS outperformed other feature selection methods (average r value = 0.81) using HV genes (average r value = 0.73) and T-DE genes (average r value = 0.63) (Figure S4E). Finally, we tested how the condition number varies as a function of the Pareto index for the results of the optimization (noise level = 0) (Figure S5). The condition number is another measure of collinearity, and features with low condition number are ideal for regression (Newman et al., 2015). We observed a




C

| | Methods | r-value | p-value | exe. time |
|---|---|---|---|---|
| unsup. | **AutoGeneS** | **0.77** | **0.05** | **168 s** |
| | HV | 0.33 | 0.38 | 7703 s |
| | T-DE | 0.56 | 0.15 | 990 s |
| | CIBERSORTx | 0.57 | 0.17 | 1 day, 22 h |
| | LM22 | 0.61 | 0.11 | 2991 s |
| | xCell | - | - | >3 days |
| | immunoStates | 0.50 | 0.16 | 934 s |
| | MCP-counter | 0.21 | 0.45 | 4140 s |
| sup. | **Shiny** | **0.78** | **0.02** | **802 s** |

**Figure 3. Evaluation of AutoGeneS for bulk deconvolution of 12 PBMC samples using bulk-sorted reference profiles**

The bulk data and the reference data are taken from Monaco et al. (2019).

(A) The flow chart depicts the underlying analysis and evaluation process of AutoGeneS when benchmarked against 7 other deconvolution approaches.

(B) For each signature matrix, the p values and r values were measured by fitting a linear regression with estimated and real proportions from flow cytometry (ground truth). The regression results using AutoGeneS are represented in Figure S6.

(C) The average r values and p values as well as the execution times are summarized in the table.

(D) Shown are pair-wise correlation and distance between the cell types using the signature matrix inferred using AutoGeneS.

high correlation between condition number and Pareto index (r=0.94) in this analysis, which shows that the optimization result with minimum correlation has the lowest or close to the lowest condition number. More details on synthetic bulk and single-cell data generation are explained in the STAR methods section.

## Deconvolution of PBMC samples with sorted bulk RNA-seq

We next deconvolved bulk RNA-seq data of PBMC samples from (Monaco et al., 2019). The bulk RNA-seq data consisted of gene expression measurements of 12 Singaporean individuals. The sorted bulk RNA samples (n = 114) of 29 immune cell types in the blood samples of four healthy individuals were also available from (Monaco et al., 2019) to build the signature matrix. The flow cytometry proportions of each sample were employed for validating the deconvolution results. The authors (Monaco et al., 2019) delineated the cell types with the highest mean Pearson correlation and merged those from the classification used for FACS (i.e., the 29 cell types) with no detectable and/or no specific signal into broader cell types. From this procedure, they selected 17 cell types that we also used in our analysis.

To highlight the advantages and potential limitations of the proposed method, we compared AutoGeneS to three other unsupervised feature selection methods, including HV genes, T-DE genes, and CIBERSORTx (Newman et al., 2019), as well as five predefined signatures proposed in the literature for deconvolving PBMC samples called Shiny (Monaco et al., 2019), LM22 (Newman et al., 2015), xCell (Aran et al., 2017), immunoStates (Vallania et al., 2018), and MCP-counter (Becht et al., 2016). Besides the sorted bulk RNA samples, Shiny used as reference the mixture samples weighted by flow cytometry proportions. Because features are predicted from the same mixture samples being used for deconvolution, we referred to Shiny as a supervised approach. This approach works when proportion data, by e.g., flow cytometry, is available, which for many studies is not. For some methods, namely LM22, xCell, immunoStates, and MCP-counter, only partial evaluation was possible because each proposed a signature matrix including some cell types that were different or missing in our reference profiles. Hence, we filtered the same reference profiles, as used by the unsupervised feature selection methods, with the signatures proposed by each of those methods. Because our goal was to study the effect of different signatures on the accuracy of the results, we also employed the same regression method, Nu-SVR, for all the signature matrices to infer the cellular proportions (Figure 3A, see STAR methods).

The linear agreement between the real proportions from flow cytometry and the predicted ones was determined by correlation coefficient (r value) and two-sided p value (Figure 3B). xCell could not perform the regression within 3 days of

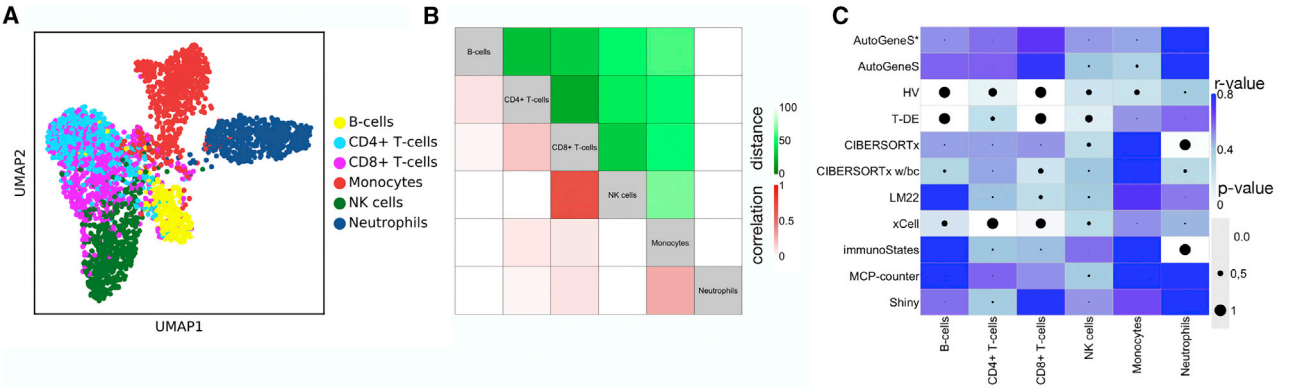



**Figure 4. Evaluation of AutoGeneS for bulk deconvolution of 12 PBMC samples using single-cell reference profiles (n = 1 samples)**

The bulk samples are taken from Monaco et al. (2019), while the single-cell reference data is taken from Reusch et al. (2021).

(A) UMAP of PBMC scRNA- seq data. Colors: six major cell types.

(B) Shown are pair-wise correlation and distance between the cell types using the signature matrix inferred using AutoGeneS.

(C) p values and r values for the deconvolution results using various signature matrices. The assessment process is similar to Figure 3A. The p values and r values were measured by fitting a linear regression with estimated and real proportions from flow cytometry. The regression results using hierarchical AutoGeneS is represented in Figure S7.

computation time; therefore, we excluded that for this comparison. Excluding Shiny, AutoGeneS had the best prediction accuracy (r value = 0.77) and other methods failed to capture the correct proportions for several cell types (r values < 0.77, Figures 3B and 3C). Moreover, although the accuracy of supervised Shiny is only slightly higher (r value = 0.78) than AutoGeneS, AutoGeneS is unsupervised and consequently has broader applications than Shiny. Results also show that AutoGeneS successfully found the signatures that minimize the correlation while simultaneously maximizing the distance between each pair of cell types (average pair-wise correlation = 0.07, Figure 3D). Besides, AutoGeneS had significantly less computation time (Figure 3C). The reduction in the computation time is mostly because AutoGeneS used only 500 genes, whereas, e.g., Shiny used 1,296.

## Deconvolution of PBMC samples with scRNA-seq

We next turned to the problem of inferring cellular proportions when bulk and single-cell reference data originate from different studies. We deconvolved PBMC samples from (Monaco et al., 2019) using the scRNA-seq data (SeqWell) of blood from Reusch et al. (2021). The single-cell data (n = 1 sample) has a total of 3,417 cells and consists of cell types including: adipocytes, B cells, $CD4^+$ T cells, CD8+ T cells, chondrocytes, dendritic cells, endothelial cells, eosinophils, erythrocytes, hematopoietic stem cells, macrophages, monocytes, natural killer (NK) cells, neutrophils, and skeletal muscle. We filtered out populations with fewer than 45 cells and were left with six major cell types. NK cells, CD4+ T cells, and CD8+ T cells have a high correlation that is also visible in the uniform manifold approximation and projection (UMAP) (McInnes et al., 2018a, 2018b) of the single-cell data (Figure 4A). Here, we assessed how the aforementioned unsupervised feature selection techniques as well as PBMC-specific signatures proposed in the literature dealt with high correlation using a similar evaluation process as in Figure 3A. After subsetting the reference profiles, the proportions were inferred for each signature matrix using Nu-SVR. We also performed CIBERSORTx with batch correction (referred to as CIBERSORTx w/bc) to minimize the technical differences between the signature matrix derived from the single-cell reference profiles and the mixture samples. The existing unsupervised methods mostly failed to infer the proportion of B cells, neutrophils, and CD4+ T cells, while AutoGeneS outperformed them. We observed that CIBERSORTx w/bc (r = 0.43) performed better than basic CIBERSORTx (r = 0.26) that suggests using batch correction after feature selection to improve deconvolution results. Also, AutoGeneS (r = 0.60) had the closest results to Shiny (r = 0.61) while even predicting the fraction of NK cells more accurately (Figure 4C). Interestingly, MCP-counter showed the highest performance (r = 0.66) for this reference data while its performance using sorted bulk reference profiles was extremely low (r = 0.21) (Figure 3).

Compared with the ground truth cell proportions as determined by flow cytometry, the low correlation in the deconvolution results might be driven either by platform-specific variation between bulk and scRNA-seq data or the inability of the feature selection techniques to extract discriminative genes. However, the lower accuracy of Shiny (r = 0.61) using the single-cell reference profiles compared to the previous results where sorted bulks were used (r = 0.78) (Figure 3B) validates the former hypothesis (for validation on synthetic data with fixed bulk samples, see Figure S4). Direct comparison of AutoGeneS and flow cytometry results is presented in Figure S7.

## Deconvolution of ethmoid sinus from patients with nasal polyposis confirms a depletion in glandular cells and enrichment in basal cells

Subsets of epithelial cells—including secretory (or apical) and ciliated cells—differentiate from basal cells to protect the upper airway (Whitsett and Alenghat, 2015), (Iwasaki et al., 2017), (Ordovas-Montanes et al., 2018). Allergic inflammation in the upper airway can result in chronic rhinosinusitis like severe nasal polyps (Ordovas-Montanes et al., 2018). During chronic rhinosinusitis, the composition of overall tissue cellular ecosystem might shift in humans. Ordovas-Montanes et al. (Ordovas-Montanes et al., 2018) revealed a remarkable reduction in cellular

diversity of polyps mostly characterized by enrichment in basal cells, concomitant loss of glandular cells, and phenotypic shifts in secretory cell expression. To confirm these findings with deconvolution results, we applied AutoGeneS to 27 mixture samples from the ethmoid sinus of patients with polyposis (n = 17) and non-polyposis (n = 10). As reference data, we used scRNA-seq data of n = 5 non-polyp and n = 6 polyp-patient ethmoid sinus. All bulk and single-cell datasets in this study are from (Ordovas-Montanes et al., 2018). A total of 10 cell types—including apical, basal, ciliated, endothelial, fibroblast, glandular, mast, myeloid, plasma, and T cells—was characterized in the reference dataset by Ordovas-Montanes et al. (Ordovas-Montanes et al., 2018) (Figure 5A). From the reference dataset, apical and basal cells are very sensitive to disease state and their signatures vary between the polyp and non-polyp samples. Therefore, we performed a data-driven clustering on the single-cell dataset and assigned state-specific annotations to apical and basal cells, called apical_non, apical_polyp, basal_non, and basal_polyp (Figure 5A, see STAR methods). 85% of new cell annotations agree with the ones used in (Ordovas-Montanes et al., 2018). This helped to find state-specific markers during feature selection that improved deconvolution (Figure 5B).

Next, we evaluated the performance of AutoGeneS, CIBERSORTx, and MuSiC in predicting the proportion of cell types in the bulk samples. Using AutoGeneS, we obtained 400 genes that differentiated the 12 cell types (Figure S8A) and successfully captured the genes associated with disease state in epithelial cells—apical, basal, ciliated, and glandular cells (Figure 5B). In our results, AutoGeneS exhibited superior performance in capturing shifts in the composition of non-polyps versus polyps mostly driven by a decrease in glandular cell proportion and an increase in basal cell proportion that are significantly concordant with the related study (Ordovas-Montanes et al., 2018) (Figures 5C, 5E, and S8C). While CIBERSORTx failed to infer the proportion of basal cell and MuSiC failed to reveal the cellular diversity of non-polyps and polyps (Figures 5C and 5E). We should also highlight that only AutoGeneS predicted the proportion of state-specific apical cell properly in non-polyps and polyps (Figure S8C). Due to the high correlation between basal_non and basal_polyp cells and the lack of enough unique biomarkers that differentiate those, their proportions are not individually conclusive (Figures 5B and S8C).

Among epithelial cell types, ciliated and glandular cells are histologically distinguishable and have non-overlapping gene biomarkers. Therefore, to further confirm our results, we characterized the association of the proportion of ciliated and glandular cells with the level of C11orf88 and CA2, two important biomarkers for ciliated and glandular cells, respectively (Figure S8D). Based on our results, AutoGeneS identified a high correlation between the proportion of ciliated and glandular cells and C11orf88 and CA2 levels (average r value = 0.84), while other methods performed substantially worse (average r value = 0.66 and 0.04 for CIBERSORTx and MuSiC, respectively, Figure 5D).

Finally, we examined the average number of molecule counts in each cell type. The single-cell dataset indicates a significant variation in average cell count across cell types mostly driven by a high number of counts in apical non, ciliated, and glandular cells, and a low number of counts in mast and T cells. This is mainly because of different cell size—the average number of mRNA molecules in each cell type—and might lead to a higher inferred proportion for larger cells (see STAR methods). With no prior knowledge of cell sizes, predicting absolute cellular proportions using conventional methods is impossible. Assuming that the average number of molecule counts in a cell type is roughly comparable with the size of that cell (Padovan-Merhar et al., 2015), (Wang et al., 2019), we applied regression on the average gene expression of cells without normalization as an alternative approach to getting cell-type-specific relative abundance rates (referred to as AutoGeneS+ in Figure S8C). Considering cell size, the absolute cellular proportions changed, however the fold change of cell-type proportions across samples is similar to previous results (see STAR methods).

***Hierarchical optimization for highly correlated cell types***

The correlation matrix reflects the similarity between cell types. Based on this, the user then groups the cells with almost no differentiating signals. After running the optimization, the correlation matrix on the selected features reveals whether some of the cell types are still correlated. If the correlation is high for some cell types, we propose running AutoGeneS at a different stage on those correlated cell types and aggregating the new features with the earlier ones to build the final signature matrix. If the number of added genes is high, the correlation between other cell types can potentially increase. To avoid that, we can run the optimization with those genes being preselected and fixed in the new solutions. Therefore, the optimization finds genes that in combination with those preselected ones optimize the objectives. For the dataset studied in Figure 4, we ran AutoGeneS separately for CD4+ and CD8+ T cells and added in total 10 more genes to the signature matrix that can better differentiate between these two cell types. For the example in Figure 4, the results show that the hierarchical AutoGeneS (AutoGeneS*) obtained better results compared with basic AutoGeneS specifically for CD8+ T cells and monocytes (Figure 4C).

## DISCUSSION

The knowledge of cell-type compositions is important to reveal cellular heterogeneity in diseased tissues and is helpful to identify the targets of a disease. Although bulk RNA-seq averages gene expression across thousands of cells and, thus, masking cellular heterogeneity in complex tissues, scRNA-seq usually does not reflect true cell-type proportions in intact tissues. Furthermore, it remains costly for use in clinical studies with large cohorts. Therefore, several bulk deconvolution techniques have been proposed in the literature to predict cell-type proportions from mixed samples. Existing approaches either rely on preselected marker genes or perform a single-criterion test to identify differentially expressed genes among cell types. In this study, we introduce AutoGeneS which, for the first time, selects discriminative genes based on multiple criteria and infers accurate cell-type proportions. AutoGeneS selects discriminative genes using a MOO approach that simultaneously minimizes average pair-wise correlation coefficients and maximizes the Euclidean distance between cell types. By minimizing correlation, we reduce the collinearity as one of the main challenges in bulk deconvolution. We simultaneously ensure that the Euclidean distance between the cell types is maximized to safeguard

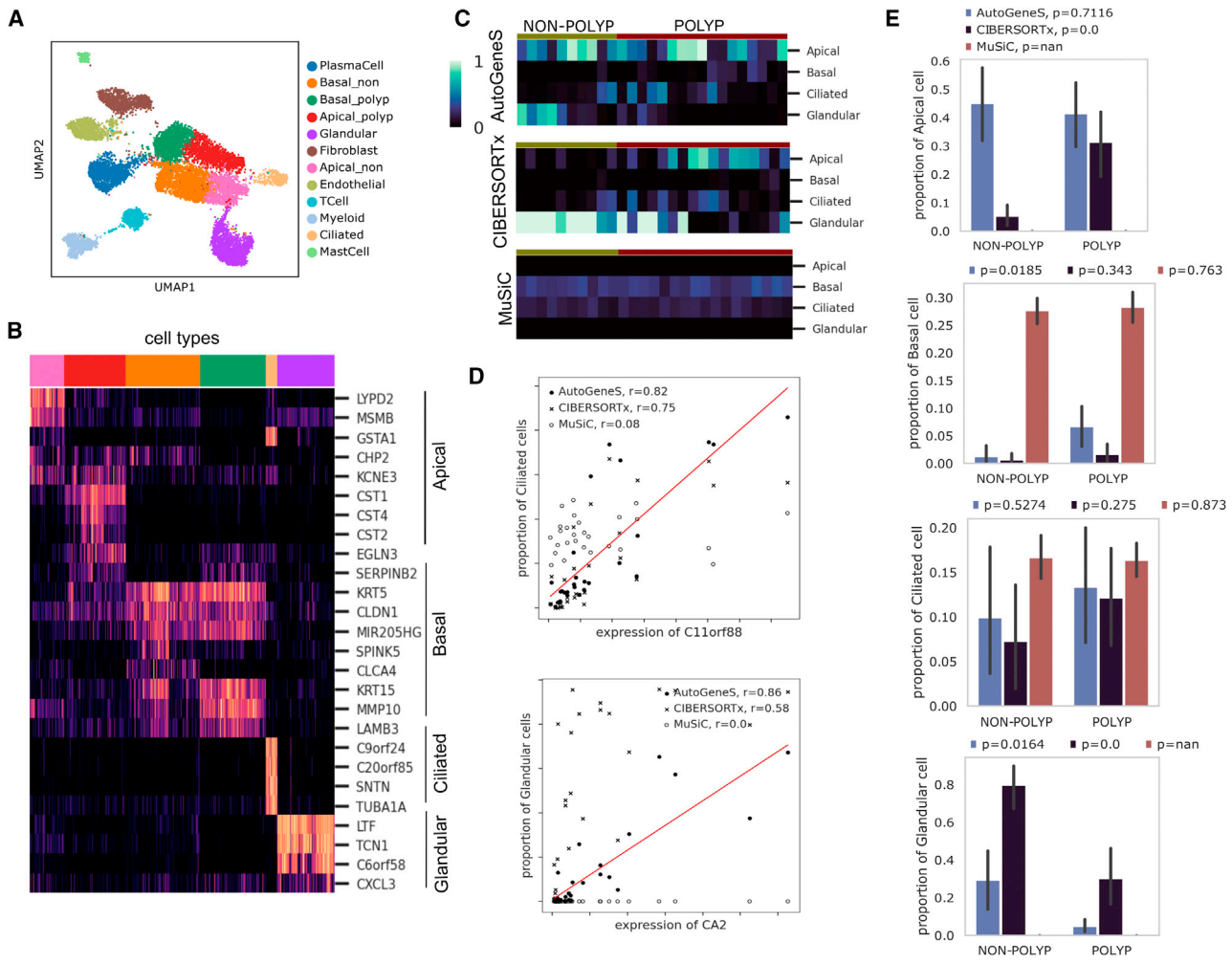



**Figure 5. Bulk deconvolution results of 27 (n = 10 non-polyp and n = 17 polyp) ethmoid sinus samples**

(A) UMAP plot of 14,878 single cells from ethmoid sinus (n = 6 non-polyp and n = 5 polyp samples) taken from Ordovas-Montanes et al. (2018). Colors highlight 12 cell types where 85% of cell annotations overlap with the annotations characterized in Ordovas-Montanes et al. (2018).

(B) The expression of signatures distinguishing apical (secretory), basal, ciliated, and glandular cells across single cells from those cell types. The signatures (n = 52) are the most differentially expressed genes between cell types selected from 400 discriminative ones assessed by AutoGeneS (see STAR methods). The expression of these genes across bulk samples is shown in Figure S8B.

(C) Proportion of apical, basal, ciliated, and glandular cells across 27 samples using AutoGeneS, CIBERSORTx, and MuSiC. For the sake of visualization, the proportion of apical_non and apical_polyp as well as basal_non and basal_polyp are summed up in the heatmap. The proportion of 12 cell types are presented in Figure S8C.

(D) Top, scatterplot depicting concordance between the expression of C11orf88 and the frequency of ciliated cell quantitated by the three methods. Bottom, similarly, concordance between the expression of CA2 and the frequency of the glandular cell. The significance of the results was assessed by r values from linear regression.

(E) Composition shift in the non-polyp and polyp cellular ecosystems of the ethmoid sinus samples. Statistical comparisons were performed using a two-sided t test, non-polyp versus polyp samples.

predictions against noise. Through comprehensive benchmark evaluations and analysis of multiple real datasets, we show that AutoGeneS outperforms other approaches. Moreover, ground truth cell-type proportions identified by flow cytometry confirm the accuracy of the predictions by AutoGeneS.

While in this study we use correlation and distance as objectives, AutoGeneS offers a flexible framework that can be easily extended to other user-specific objectives. Among the unsupervised feature selection techniques previously proposed for bulk deconvolution, we implemented an objective function inspired by the works in (Newman et al., 2015). Newman et al. (Newman et al., 2015) proposed first preselecting top G marker genes for each population using a t test with a low q value and later iterating G across all populations and retraining the signature matrix with the lowest condition number. We studied a new implementation of this approach by integrating condition number into AutoGeneS as an objective function to be minimized. We compared that with other single-objective optimizations using correlation and distance as well as the MOO for the dataset underlying Figure 3 (Figures S10 and S11). We observed that the

MOO (max r value = 0.77 and average std = 0.014) outperformed the single-objective optimization using condition (max r value = 0.69 and average std = 0.033), distance (max r value = 0.29 and average std = 0.013), or correlation (max r value = 0.76 and average std = 0.031). Plotting the objective values of the optimal solution as a function of r and p values, we observed that minimizing condition number does not necessarily improve the predictions (the minimum condition number using 200 genes is lower than 1,200 genes while the predictions using 1,200 genes are more accurate). However, correlation is negatively associated with r value. Therefore, correlation can be employed as a measure to select the right number of genes for deconvolution.

AutoGeneS currently requires reference profiles of desired cell types to search for discriminative genes. The reference profiles can be provided by either sorted bulk RNA-seq or scRNA-seq. Using scRNA-seq data has the advantage of putatively revealing novel cell types and subtypes in a system that no prior knowledge of marker genes for such novel populations exists. We have demonstrated in this work that AutoGeneS can successfully utilize scRNA-seq data as effective references to identify marker genes for differentiating cell types.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Simulated bulk RNA sample generation
  - Simulated scRNA-seq data generation
  - Deconvolution
  - Selection of a solution from Pareto-optimal solutions
  - Highly variable genes
  - Silhouette score
  - Regression
  - Normalization
  - Benchmarks
  - Visualisation

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2021.05.006.

## ACKNOWLEDGMENTS

We would like to thank the many people who proofread the case study notebook and the manuscript and improved it with their comments and expertise. For this, we acknowledge the input of C. Marr, A. Wojtuszkiewicz, B. Schubert, M. Heinig, S. Palit, I. Ibarra, M. Luecken, and L. Zappia. We would like to also thank J. Ordovas-Montanes for his valuable comments on the results for ethmoid sinus. This work was supported by the Helmholtz Association's Initiative and Networking Fund through Helmholtz AI (grant number: ZT-I-PF-5-01) and sparse2big (grant number ZT-I-007). F.J.T. reports receiving consulting fees from Roche Diagnostics and Cellarity and an ownership interest in Cellarity and Dermagnostix.

## AUTHOR CONTRIBUTIONS

Conceptualization and methodology, H.A. and F.J.T.; software and analysis, H.A.; writing, H.A. and F.J.T., funding acquisition, F.J.T.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 8, 2020
Revised: July 31, 2020
Accepted: May 7, 2021
Published: June 7, 2021

## REFERENCES

Aran, D., Hu, Z., and Butte, A.J. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. *18*, 220.

Baron, M., Veres, A., Wolock, S.L., Faust, A.L., Gaujoux, R., Vetere, A., Ryu, J.H., Wagner, B.K., Shen-Orr, S.S., Klein, A.M., et al. (2016). A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. Cell Syst *3*, 346–360.e4.

Becht, E., Giraldo, N.A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautès-Fridman, C., Fridman, W.H., and de Reyniès, A. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. *17*, 218.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) *57*, 289–300.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech.: Theor. Exp *10*, 10008.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., and Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods *10*, 1093–1095.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Büttner, M., Miao, Z., Wolf, F.A., Teichmann, S.A., and Theis, F.J. (2019). A test metric for assessing single-cell RNA-seq batch correction. Nat. Methods *16*, 43–49.

Chang, C.-C., and Lin, C.-J. (2011). Libsvm: A library for support vector machines. ACM Trans. Intell. Syst. Technol. *2*, 1–27.

Chen, P., Lin, C., and Schölkopf, B. (2005). A tutorial on $\nu$ -support vector machines. Appl. Stochastic Models Bus. Ind. *21*, 111–136.

Chu, L.F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D.T., Choi, J., Kendziorski, C., Stewart, R., and Thomson, J.A. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biol. *17*, 173.

De Rainville, F.-M., Fortin, F.-A., Gardner, M.-A., Parizeau, M., and Gagne, C. (2012). DEAP: a python framework for evolutionary algorithms. In Proceedings of the 14th Annual Conference Companion on Genetic and Evolutionary Computation, p. 8592. https://doi.org/10.1145/2330784.2330799.

Deb, K. (2011). Multi-objective optimisation using evolutionary algorithms: an introduction. In Multi-objective Evolutionary Optimisation for Product Design and Manufacturing (Springer), pp. 3–34.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic. IEEE Transactions on Evolutionary Computation *6*, 182–197.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., Marquez, J.R., Gruber, B., Lafourcade, B., Leitao, P.J., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography *36*, 27–46.

Du, R., Carey, V., and Weiss, S.T. (2019). deconvSeq: deconvolution of cell mixture distribution in sequencing data. Bioinformatics *35*, 5095–5102.

Enge, M., Arda, H.E., Mignardi, M., Beausang, J., Bottino, R., Kim, S.K., and Quake, S.R. (2017). Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. Cell *171*, 321–330.e14.

Frishberg, A., Peshes-Yaloz, N., Cohn, O., Rosentul, D., Steuerman, Y., Valadarsky, L., Yankovitz, G., Mandelboim, M., Iraqi, F.A., Amit, I., et al. (2019). Cell composition analysis of bulk genomics using single-cell data. Nat. Methods *16*, 327–332.

García-Mart, Í.C., Rodriguez, F.J., and Lozano, M. (2018). Genetic algorithms. In Handbook of Heuristics (Springer International Publishing), pp. 431–464.

Görtler, G.F., Solbrig, S., Wettig, T., Oefner, P.J., Spang, R., and Altenbuchinger, M. (2018). Loss- function learning for digital tissue deconvolution. J Comput Biol. *27*, 342–355.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849. https://doi.org/10.1093/bioinformatics/btw313.

Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., et al. (2020). Construction of a human cell landscape at single-cell level. Nature *581*, 303–309.

Hunt, G.J., Freytag, S., Bahlo, M., and Gagnon-Bartsch, J.A. (2019). Dtangle: accurate and robust cell type deconvolution. Bioinformatics *35*, 2093–2099.

Iwasaki, A., Foxman, E.F., and Molony, R.D. (2017). Early local immune defences in the respiratory tract. Nat. Rev. Immunol *17*, 7–20.

Kang, K., Meng, Q., Shats, I., Umbach, D.M., Li, M., Li, Y., Li, X., and Li, L. (2018). A novel computational complete deconvolution method using RNA-seq data. bioRxiv https://www.biorxiv.org/content/10.1101/496596v1.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Konak, A., Coit, D.W., and Smith, A.E. (2006). Multi-objective optimization using genetic algorithms: a tutorial. Reliab. Eng. Syst. Saf *91*, 992–1007.

Kuhn, A., Kumar, A., Beilina, A., Dillman, A., Cookson, M.R., and Singleton, A.B. (2012). Cell population-specific expression analysis of human cerebellum. BMC Genomics *13*, 610.

Kuhn, A., Thu, D., Waldvogel, H.J., Faull, R.L., and Luthi-Carter, R. (2011). Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. Nat. Methods *8*, 945–947.

McInnes, L., Healy, J., and Melville, J. (2018a). UMAP: uniform manifold approximation and projection for dimension reduction. aRxiv http://arxiv.org/abs/1802.03426.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018b). Umap: uniform manifold approximation and projection. J. Open Source Software *3*, 861. https://doi.org/10.21105/joss.00861.

Monaco, G., Lee, B., Xu, W., Mustafah, S., Hwang, Y.Y., Carr e, C., Burdin, N., Visan, L., Ceccarelli, M., Poidinger, M., et al. (2019). RNA-seq signatures normalized by mrna abundance allow absolute deconvolution of human immune cell types. Cell Rep *26*, 1627–1640.e7.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.

Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Kho- dadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol *37*, 773–782.

Ordovas-Montanes, J., Dwyer, D.F., Nyquist, S.K., Buchheit, K.M., Vukovic, M., Deb, C., Wadsworth, M.H., Hughes, T.K., Kazer, S.W., Yoshimoto, E., et al. (2018). Allergic inflammatory memory in human respiratory epithelial progenitor cells. Nature *560*, 649–654.

Padovan-Merhar, O., Nair, G.P., Biaesch, A.G., Mayer, A., Scarfone, S., Foley, S.W., Wu, A.R., Churchman, L.S., Singh, A., and Raj, A. (2015). Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. Mol. Cell *58*, 339–352.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Reusch, N., Bassler, K., and Schultze, J. (2021). https://beta.fastgenomics.org/datasets/ detail-dataset-4ed9c26de21346268965537b892ec25b.

Schelker, M., Feau, S., Du, J., Ranu, N., Klipp, E., Macbeath, G., Schoeberl, B., and Raue, A. (2017). Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nat. Commun *8*, 2032.

Schwartz, R., and Shackney, S.E. (2010). Applying unmixing to gene expression data for tumor phylogeny inference. BMC Bioinformatics *11*, 42.

Segerstolpe, Å., Palasantza, A., Eliasson, P., Andersson, E., Andréasson, A., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M.K., et al. (2016). Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell Metab *24*, 593–607.

Shen-Orr, S.S., and Gaujoux, R. (2013). Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr. Opin. Immunol *25*, 571–578.

Shen-Orr, S.S., Tibshirani, R., Khatri, P., Bodian, D.L., Staedtler, F., Perry, N.M., Hastie, T., Sarwal, M.M., Davis, M.M., and Butte, A.J. (2010). Cell type-specific gene expression differences in complex tissues. Nat. Methods *7*, 287–289.

Smith, A.E. (2005). Multi-objective optimization using evolutionary algorithms [Book Review]. IEEE Trans. Evol. Computat *6*, 526.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21.

Tran, H.T.N., Ang, K.S., Chevrier, M., Zhang, X., Lee, N.Y.S., Goh, M., and Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. *21*, 12.

Vallania, F., Tam, A., Lofgren, S., Schaffert, S., Azad, T.D., Bongen, E., Haynes, W., Alsup, M., Alonso, M., Davis, M., et al. (2018). Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. Nat. Commun. *9*.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., , et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in python. Nat Methods *17*, 261–272.

Wang, X., Park, J., Susztak, K., Zhang, N.R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat. Commun *10*, 380.

Whitsett, J.A., and Alenghat, T. (2015). Respiratory epithelial cells orchestrate pulmonary innate immunity. Nat. Immunol *16*, 27–35.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

Zaitsev, K., Bambouskova, M., Swain, A., and Artyomov, M.N. (2019). Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. Nat. Commun *10*, 2209.

# STAR★METHODS

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Fabian Theis (fabian.theis@helmholtz-muenchen.de).

### Materials availability

This study did not generate new materials.

### Data and code availability

- Source data statement: In this study, we used existing datasets which are openly available as cited in the references section. The single-cell source datasets are listed in Table S1. The bulk RNA samples, as well as sorted bulk samples underlying Figures 3, 4 and S6 are downloaded from GEO with accession codes GEO: GSE106898 and GEO: GSE10701, respectively. Finally, the bulk RNA samples underlying Figures 5 and S8 are accessible in Table S11 of Ordovas-Montanes et al. (Ordovas-Montanes et al., 2018).
- Code statement: AutoGeneS is publicly available as a Python package at Github: https://github.com/theislab/AutoGeneS. A tutorial and an example are also provided.
- Scripts statement: The scripts used to generate the figures reported in this paper are available at Github: https://github.com/theislab/reproducibility-AutoGeneS.
- Any additional information required to reproduce this work is available from the lead contact, Fabian Theis (fabian.theis@helmholtz-muenchen.de).

## METHOD DETAILS

### Simulated bulk RNA sample generation

To validate the proposed deconvolution process, we required bulk RNA-seq with known cell-type proportions. The single-cell reference data underlying Figure 2 has six cell types for which matched sorted bulk RNA samples (n=16 with at least 2 replicas per cell type) are also available. We generated in total of 100 synthetic bulk profiles as a mixture of the sorted bulk RNA samples. For each synthetic bulk, we subsampled from the sorted bulks to have at least one sample per cell type and then multiplied the gene expression profiles with a random fraction (sum of fractions=1.0). We used the fractions later to validate the inferred proportions using our method. Because the single-cell reference profiles and the synthetic bulk RNA samples were generated using different sequencing technologies, this was a more realistic way of generating synthetic bulk RNA samples compared with the approaches that subsample scRNA-seq data and sum the expression profiles to simulate bulk RNA samples.

### Simulated scRNA-seq data generation

In the scRNA-seq data underlying Figure 2, the cell types are well defined with low correlation. To evaluate our method concerning different noise levels as well as higher cellular correlation, we modified the single-cell data as follows: We first removed top differentially expressed genes (n=5,379) from a total of 19,089 genes. We then removed highly variable genes (n=5,000) from the remaining ones and a total of 8,719 genes was left for efficiency analysis. Since the gene selection in AutoGeneS is based on the mean expression of cells, dropout— low mRNA detection in single-cell data—can influence the efficiency of AutoGeneS. Therefore, we added to the noise level through dropping out genes from the single-cell data and re-generated the signature matrix for each noise level. Depending on the dropout rate, genes are randomly selected and set to 0. We tested various dropout rates between 0−50 percent (rate=r means that r percent of genes in each cell were randomly set to 0). The results in Figure S4 show that AutoGeneS, compared to highly variable (HV) and differentially expressed (T-DE) genes, is more robust to this source of noise. However, increasing the dropout rate decreases r-value as expected.

### Deconvolution

To infer cellular proportion, related work generally assumes that the count matrix Y of m mixture samples with N genes is the weighted sum of K cell-type-specific count profiles with the same N genes, represented by matrix X (Kuhn et al., 2011), (Frishberg et al., 2019), (Görtler et al., 2018). This can be modeled as a system of linear equations as follows:

$$Y = XW + E \quad \text{(Equation 1)}$$

where, W is an k × m fractional abundance matrix and E represents the residual errors. To compute W, we solve the equation for only a subset of genes g << N selected using AutoGeneS:

$$Y^g = X^g W + E \quad \text{(Equation 2)}$$

RNA count data is often normalized for library size to remove any difference that is arisen due to sampling effects. However, this approach can mask the difference between biological samples with different mRNA abundance. To consider the total mRNA abundance of each cell type, we can add the mean library size of cell types to Equation 2 as follows:

$$Y^g = X^g S W' + E \quad \text{(Equation 3)}$$

where, S is a k × k diagonal matrix with $S_{ll}$ being the average number of mRNAs in cell type l (also called cell size). Considering S in the regression function, absolute cellular proportions canchange ($W \neq W'$), however for a cell-type specific comparison across samples i and j, $\frac{W_{l,i}}{W_{l,j}}$ is still equal to $\frac{W'_{l,i}}{W'_{l,j}}$.

To build the signature matrix using bulk-sorted RNA-seq, we averaged transcripts-per-million (TPM)-normalized cell-type-specific profiles in non-log linear space and filtered the markers. Later, we used the mean expression of all cells in a population after normalizing each cell into TPM and filtered the markers to obtain population-specific reference profiles. Multi-objective optimization. The proposed feature selection approach solves a multi-objective optimization problem that can be generally formulated as: subject to:

$$\text{minimize/maximize } \vec{f}(\vec{z}) := [f_1(\vec{z}), ....., f_k(\vec{z})] \quad \text{(Equation 4)}$$

$$h_i(\vec{z}) \leq 0, i = 1, ..., m \quad \text{(Equation 5)}$$

$$e_i(\vec{z}) = 0, i = 1, ..., p \quad \text{(Equation 6)}$$

where $\vec{z}$ is the vector of decision variables and $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$, i = 1,..., k are the objective functions which evaluate the quality of a solution by assigning a fitness value to it. Also, $h_i, e_j : \mathbb{R}\ n \rightarrow \mathbb{R}$, i = 1,...,m, j = 1,...,p are referred to as the inequality and equality constraint functions of a problem which must be satisfied. In AutoGeneS, we have n binary decision variables where n is equal to the number of genes from which the optimizer selects the markers. The value of a decision variable represents whether the corresponding gene is selected as a marker. Later, we evaluate the objective functions (correlation and distance) only for genes G whose decision variables are set to one. Considering $X_k^G$ as the expression profile of cell type k for genes G, the correlation objective is measured as:

$$C^G = \sum_{\substack{\forall\ k \in K \\ \forall k' \in K,\ k \neq k'}} \left| \frac{X_k^G X_{k'}^G}{\|X_k^G\| \|X_{k'}^G\|} \right| \quad \text{(Equation 7)}$$

and the distance objective is measured as:

$$D^G = \sum_{\substack{\forall\ k \in K \\ \forall k' \in K,\ k \neq k'}} \|X_k^G - X_{k'}^G\| \quad \text{(Equation 8)}$$

AutoGeneS allows setting the number of selected markers G as a fixed value that can be implemented as a constraint (all the solutions where |G| is larger than a desired value, are marked as infeasible and will not be evaluated). However, we show later how this constraint is implemented differently in AutoGeneS to obtain a better performance.

Several techniques have been proposed in the literature for solving multi-objective optimization problems. Among these, metaheuristics such as multi-objective evolutionary algorithms (MOEAs) are quite popular and well established mainly because of their flexibility and widespread applications (Deb et al., 2002), (Deb, 2011), (Smith, 2005). MOEA denotes a class of search methods where the decisions are made in the presence of trade-offs between objectives. As the name suggests, multi-objective optimization involves more than one objective function to be optimized at once. When no single solution exists that simultaneously optimizes each objective, the objective functions are said to be conflicting. In this case, the optimal solution of one objective function is different from that of the others. This gives rise to a set of trade-off optimal solutions popularly known as Pareto-optimal solutions. The list of Pareto-optimal solutions includes non-dominated solutions, which are explored so far by the search algorithm. These solutions cannot be improved for any of the objectives without degrading at least one of the other objectives. Without additional subjective preference, all Pareto-optimal solutions are considered to be equally good.

AutoGeneS uses a genetic algorithm (GA) as one of the main representatives of the family of MOEAs. GA uses a population-based approach where candidate solutions that represent individuals of a population are iteratively modified using heuristic rules to increase their fitness (i.e., objective function values) (García-Mart et al., 2018), (Konak et al., 2006). The main steps of the generic algorithm are as follows:

1. Initialization step: Here the initial population of individuals is randomly generated. Each individual represents a candidate solution that, in the feature selection problem, is a set of marker genes. The solution is represented as a bit string with each bit representing a gene. If a bit is one, the corresponding gene is selected as a marker.

2. Evaluation and selection step: Here the individuals are evaluated for fitness (objective) values, and they are ranked from best to worst based on their fitness values. After evaluation, the best feasible individuals are then stored in an archive according to their objective values.
3. Termination step: Here, if the termination conditions (e. g., if the simulation has run a certain number of generations) are satisfied, then the simulation exits with the current solutions in the archive. Otherwise, a new generation is created.
4. If the simulation continues, the next step is creating offspring (new individuals): The general GA modifies solutions in the archive and creates offspring through random-based crossover and mutation operators. First, parents are selected among the candidates in the archive. Second, the crossover operator combines the bits of the parents to create the offspring. Third, the mutation operator makes random changes to the offspring. Offspring are then added to the population, and the GA continues with step 2.

We modified the python package DEAP (v1.0 https://github.com/DEAP/deap) (De Rainville et al., 2012) to adapt it to our problem. The algorithms in Figure S9 represent how the crossover and mutation operators are implemented in AutoGeneS to always return individuals with a fixed number of marker genes. In our feature selection problem, we are generally interested in finding the most discriminative genes out of N variable genes (e. g., (Baron et al., 2016),000 HV genes). This results in a search space with $2^{|N|}$ possible solutions. However, the number of possible solutions is decreased to $\binom{|N|}{|G|}$ when we only search for solutions with a fixed number of genes. Therefore, this implementation avoids generating infeasible solutions (where the number of markers are not equal to |G|) and significantly improves the efficiency of the whole GA. For all the studies in this paper, we showed that $|G| \in [200 - 500]$ provided sufficiently good results. The number of discriminative genes depends on the number of cell types and their cellular correlation. For the desired solution, the correlation matrix showing correlation coefficients between cell types is a good assessment of the number of genes. In case the cellular correlation is still high after optimization, either more genes should be added or very similar cell types (usually with correlation> 90%) should be merged as one cell type. The number of genes also influences the execution time of the objective function (Figure S4C).

It should be noted that the number of HV genes affects the efficiency of AutoGeneS because many differentially expressed genes might not appear in the top HV genes (e. g., $n < 1, 000$). On the other side, selecting a large number of HV genes (e. g., $n > 10, 000$) increases the search space and slows the convergence of optimization results. Feeding the optimization with 4,000-6,000 HV genes often works well. For highly correlated cell types (correlation 60%), we recommend using at least 5,000 number of HV genes.

Finally, the most time-consuming part of an optimization process is the evaluation part where the objective functions are executed. Therefore, we designed the optimizer to cache the explored solutions and their objective values to avoid re-calculating the objective functions if some of those are generated in later iterations. For a fixed number of iterations, the number of HV genes should not change the execution time as it only changes the size of the search space. However, for a low number of HV genes ($n < 2, 000$), the probability of generating similar solutions increases resulting in lower execution time (Figure S4C). When the number of HV genes increases ($n > 2, 000$), the execution time slightly escalates.

### Selection of a solution from Pareto-optimal solutions

By their nature, multi-objective optimization problems give rise to a set of Pareto-optimal solutions that need further processing to find a single solution that satisfies the subjective preferences of the users. AutoGeneS also plots the set of Pareto-optimal solutions; comparing the correlation and the distance of the two extremes (the solutions with lowest correlation and highest distance) indicates which solution might be better. For the analysis in this paper, we selected the solution with minimum correlation for deconvolution. Regarding our results with ground truth, that solution always returned the highest r-value (Figure S4D). However, still, other solutions on the Pareto front might also work well.

### Highly variable genes

Highly variable (HV) genes are those that are informative of the variability in a dataset (Brennecke et al., 2013). These genes are usually filtered as the first step of feature selection. Depending on the complexity of a dataset, typically between 1,000 and 5,000 HV genes are selected for downstream analysis. The authors in (Klein et al., 2015) suggested that downstream analysis is only as robust as the exact choice of the number of HV genes. Due to the possible batch effects especially regarding varying library preparation protocols between reference and mixture samples, we prefer to select a higher number of HVGs, like 4,000. We used the method implemented in single-cell analysis in Python (SCANPY) (Wolf et al., 2018) for selecting HV genes in which genes are binned by their mean expression and those with the highest variance-to-mean ratio are selected as HV genes in each bin. Of course, before applying this method, genes with a low average expression are excluded (if not already done for the external datasets) as a quality control filter. Differentially expressed genes. To find differentially expressed genes in a population of interest, statistical tests like Wilcoxon rank-sum test or t test are usually used to rank genes by their difference in expression between two groups (the population of interest and the remaining populations). In the present study, we used t test from the Seurat package (Butler et al., 2018), (Stuart et al., 2019) implemented in SCANPY (tl.rank_genes_groups) with Benjamini-Hochberg (BH) corrected p-values (Benjamini and Hochberg, 1995) (called q-values) on log-transformed data. The top G marker genes with lowest q-values are selected for each population and aggregated as T-DE. We have selected G so that T-DE contained in total the same number of genes as used in AutoGeneS.

CellPress
OPEN ACCESS

### Silhouette score

Silhouette weight (SW) measures how similar a cell is to the cells within its own cluster compared to other clusters. SW ranges from −1 to +1, where a high value indicates that the cell is well matched to its own cluster and poorly matched to neighboring clusters. If the average of Silhouette weights (ASW) over all cells is high, the clusters are dense and well-separated. An ASW of 0 or -1 represents overlapping clusters. We used the implementation in scIB (https://github.com/theislab/scib/blob/master/scIB/metrics.py) to measure how batches are mixed. scIB scales ASW ($1 - |\text{ASW}|$) that we refer to it as Silhouette score. A Silhouette score of 1 represents separated clusters and mixed batches which is ideal.

### Regression

The proposed feature selection strategy is orthogonal to the regression technique used for inferring cellular proportions. AutoGeneS offers a set of regression methods including linear regression, Nu-support vector machine (Nu-SVR), and non-negative least squares (NNLS). For the analysis in this paper, we used Nu-SVR (sklearn.svm.NuSVR) (Pedregosa et al., 2011). Compared with regular linear regression, which attempts to minimize regression error, SVR tries to fit the regression error within a certain threshold. The $\varepsilon$-loss function in SVR ignores errors that are within $\varepsilon$ distance of the observed value by treating them as equal to zero. For other points outside this boundary, the loss is measured based on the distance between observed value y and the $\varepsilon$ boundary. This produces a better-fitting model by highlighting the loss of outliers alone. In Nu-SVR, $\text{Nu} \in (0, 1]$ indicates an upper bound on the fraction of training errors (poorly predicted samples) and a lower bound on the fraction of support vectors to use. The regression is performed on normalized, non-log-transformed data. After the regression, we ensure that the weights are positive and they sum to one. Thus, we set the negative values to 0 and normalize the rest to sum to one.

The execution times underlying Figures 3C and S4C were measured on an Intel(R) Xeon(R) Gold 6126 CPU at 2.60GHz with 395GB of RAM. Statistical test. Linear concordance between estimated and real proportions were assessed using linear regression (scipy.stats.linregress) (Virtanen et al., 2020). Pearson correlation (r-value) and two-sided p-value using Wald Test were determined for each test.

### Normalization

The external RNA-seq datasets were downloaded and analyzed using the authors' normalization consisting of TPM or reads per kilobase of transcript per million (RPKM). The count matrix for scRNA-seq datasets were normalized to one million counts per cell.

Cell type annotation of ethmoid sinus. Clustering was performed on both non-polyps (n=5) and polyps (n=6) single-cell samples (a total of n=14,878 cells) from (Ordovas-Montanes et al., 2018). A single-cell neighborhood graph (kNN-graph) was computed on the first 50 principal com- ponents using 30 neighbors. For clustering and cell-type annotation, Louvain-based clustering (Blondel et al., 2008) was used as implemented in louvain-igraph (v0.6.1 https://github.com/vtraag/louvain- igraph) and adopted by SCANPY (tl.louvain) with resolution parameter set to 0.9. Clusters were annotated on the basis of similarity to the original annotations (Ordovas-Montanes et al., 2018). 85% of new cell annotations overlapped with the original ones. We used state-specific annotations for apical and basal cells: in non-polyps, they were annotated as apical non and basal non, respectively. Similarly, apical and basal cells in polyps were annotated as apical polyp and basal polyp, respectively.

### Benchmarks

We used CIBERSORTx and MusiC with default parameters. CIBERSORTx recommended to perform batch correction to minimize batch effects as a source of confounding technical variation between reference and mixture samples. Batch correction is applied after feature selection. Since our main objective in this paper is to study the importance of feature selection for bulk deconvolution, batch correction was not performed on bulk samples using CIBERSORTx except for the analysis underlying Figure 4 (batch mode = S). This was necessary to make the comparison with other methods fair as the proposed batch correction can be applied to the output of all the methods. For xCell, immunoStates, MCP-counter, LM22, and Shiny, we downloaded their proposed signatures and filtered the reference profiles used in this paper for those signatures, and performed deconvolution using Nu-SVR. For the results underlying Figure 5, CIBERSORTx was performed using its built-in regression.

### Visualisation

The UMAP plots in this paper are generated using SCANPY (tl.umap) (McInnes et al., 2018a, 2018b). The heatmaps underlying Figures 2, 3, and 4 are generated using ComplexHeatmap (Gu et al., 2016).