CDSeqR: fast complete deconvolution for gene expression data from bulk tissues

Kai Kang [1,*,†], Caizhi David Huang [1,†], Yuanyuan Li [1], David M. Umbach[1], and Leping Li [1,*]

[1]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, 27709, USA

[*]To whom correspondence should be addressed.

[†]These authors contribute equally.

Emails of corresponding authors: kangkai0714@gmail.com; li3@niehs.nih.gov

Abstract

Background: Biological tissues consist of heterogenous populations of cells. Because gene expression patterns from bulk tissue samples reflect the contributions from all cells in the tissue, understanding the contribution of individual cell types to the overall gene expression in the tissue is fundamentally important. We recently developed a computational method, CDSeq, that can simultaneously estimate both sample-specific cell-type proportions and cell-type-specific gene expression profiles using only bulk RNA-Seq counts from multiple samples. Here we present an R implementation of CDSeq (CDSeqR) with significant performance improvement over the original implementation in MATLAB and an added new function to aid cell type annotation. The R package would be of interest for the broader R community.

Result: We developed a novel strategy to substantially improve computational efficiency in both speed and memory usage. In addition, we designed and implemented a new function for annotating the CDSeq estimated cell types using single-cell RNA sequencing (scRNA-seq) data. This function allows users to readily interpret and visualize the CDSeq estimated cell types. In addition, this new function further allows the users to annotate CDSeq-estimated cell types using marker genes. We carried out additional validations of the CDSeqR software using synthetic, real cell mixtures, and real bulk RNA-seq data from the Cancer Genome Atlas (TCGA) and The Genotype-Tissue Expression (GTEx) project.

Conclusions: The existing bulk RNA-seq repositories, such as TCGA and GTEx, provide enormous resources for better understanding changes in transcriptomics and human diseases. They are

also potentially useful for studying cell-cell interactions in the tissue microenvironment. Bulk level analyses neglect tissue heterogeneity, however, and hinder investigation of a cell-type-specific expression. The CDSeqR package may aid *in silico* dissection of bulk expression data, enabling researchers to recover cell-type-specific information.

Keywords: Deconvolution, tissue heterogeneity, gene expression, CDSeq.

Background

Many biological samples present heterogeneity; namely, their cellular composition varies phenotypically and functionally. The cell-type proportions can exhibit high variability cross samples from the same tissue type, potentially indicating the occurrence of various physiological and pathological activities in the samples. Therefore, knowing the cellular composition of heterogeneous samples is important to better understand the roles of distinct cell populations. For example, dissection of tumor-infiltrating immune cells holds promise for gaining deeper insights into the underlying mechanisms of cancers, for discovering predictive biomarkers, as well as for the development of novel treatment strategies [1,2]. Although single-cell RNA sequencing has become popular, in addition to its high cost and sparse read counts, isolating cells from solid tissues especially frozen tissues remains challenging. Moreover, the inevitable sparsity of RNA molecules sequenced from frozen samples further challenges computational techniques. Therefore, bulk RNA-seq data are still routinely generated in many research projects. The heterogeneous nature of bulk RNA-seq, however, hampers efforts to study how individual cell types in the tissue microenvironment influence the underlying biology

of interest. Computational deconvolution methods [3,4,13–16,5–12] are complementary to single-cell analysis and are very useful tools for extracting cell-type-specific signals from bulk measurements.

This work describes an R package (CDSeqR) implementation of CDSeq (complete deconvolution using sequencing data) [5] that we recently developed in MATLAB. CDSeq was designed to simultaneously estimate the cell-type-specific gene expression profiles (GEPs) and the sample-specific cell-type proportions using bulk RNA-seq data without the need of reference GEPs (i.e. expression of pure cell lines or scRNA-seq data). Here, we not only implemented a user-friendly R version for CDSeq but also substantially improved its computational efficiency in both speed and memory usage. Specially, we designed a dimension Reduction-Recover approach that is automatically parallelized to facilitate computation. In addition, we implemented a new functionality for annotating the CDSeq-estimated cell types using scRNA-seq data. Moreover, we employed a recently published deconvolution benchmarking pipeline [17] to thoroughly test CDSeqR using both cell mixtures [18] and simulated data [19–21]. Throughout the manuscript, we refer to “CDSeqR” as the package and “CDSeq” as the deconvolution procedure.

Implementation

Statistical model

CDSeq employs a hierarchical Bayesian modeling strategy to account for the mixing process of observed RNA reads from different cell types. Specifically, the gene expression of any particular cell type, in the form of read counts, is modeled by a Dirichlet multinomial random variable and

67 the bulk measure is therefore a weighted sum of all cell types (i.e., Dirichlet multinomial
68 random variables) constituting the bulk tissue. Then, CDSeq uses a Gibbs sampler [5,22] for
69 parameter estimation. A detailed description of the model and solver can be found in [5].
70
71 **A dimension reduce-recover strategy for fast deconvolution**
72 A practical challenge for our original MATLAB implementation of CDSeq was the high
73 computational demand, especially for large datasets. The Gibbs sampler employed for
74 parameter estimation requires storage of several large vectors whose sizes are data-dependent;
75 therefore, it can be computationally demanding especially for large datasets. We propose a
76 novel strategy for alleviating this issue. Our strategy, named dimension Reduce-Recover,
77 consists of two steps: dimension reduction and dimension recovery. This approach was based
78 on the following assumption: cell proportions in each sample are invariant with respect to the
79 genes used in deconvolution. Specifically, for a bulk RNA-seq sample, the proportions of cell
80 types can be estimated using multiple subsets of randomly selected genes. Therefore, for a bulk
81 RNA-Seq dataset, we should be able to obtain the estimated cell-type proportions for each
82 sample in the dataset regardless of the subset of genes considered. This idea allows us to
83 calculate multiple independent estimates of cell-type proportions using many subsets of
84 randomly selected genes processed in parallel. The final estimate of the cell-type proportions is
85 the average of the independent estimates. We then estimate the cell-type specific gene
86 expression profiles by post-multiplying the input data matrix by a generalized inverse of the
87 matrix of final estimated cell-type proportions (Supplementary Methods). We have carried out
88 thorough testing and demonstrated that this approach substantially speeds up computation

89 without sacrificing the performance (see Supplementary Note for details, and Supplementary
90 Fig. S9). We further evaluated the effects of the size of the gene list and the block number on
91 the performance when using the Reduce-Recovery strategy (Supplementary Fig. S9).
92
93 **Cell type assignment function in CDSeqR**
94 Bulk tissue RNA-seq samples usually present a great degree of heterogeneity. For example,
95 cancer cells in a tumor exhibit a spectrum of phenotypic and transcriptomic stages [23]. This
96 spectrum makes the assignment of identified cell types to known cell types challenging. In our
97 original MATLAB version, the cell-type assignment was based on correlation analysis using a
98 user-provided pure cell line GEPs. While this option remains available, the increasing number of
99 publicly available single cell data sets, e.g. [21,24]., provide alternative GEPs for cell-type
100 assignment. Thus, in CDSeqR, we designed and implemented a new function to perform cell-
101 type assignment for the downstream processing of CDSeq estimated cell-type-specific GEPs. In
102 addition, there may be cases where reliable reference datasets from either pure cell line or
103 single cell RNA-seq may not be available. Annotation of CDSeq estimated cell types using cell
104 type marker genes can also be performed in CDSeqR.
105
106 Specifically, this new function, *cellTypeAssignSCRNA*, embodies three approaches to annotate
107 the CDSeq-estimated cell types. First, it employs correlation analysis to measure the similarity
108 between the CDSeq-estimated cell-type-specific GEPs and the summarized cell-type specific
109 gene expression profiles from scRNA-seq data. When scRNA-seq reference data are provided,
110 *cellTypeAssignSCRNA* first sums up the read counts of all cells from the same cell type to form

111 the aggregated cell-type specific gene expression profiles for each cell type in the scRNA-seq
112 data. Next, *cellTypeAssignSCRNA* performs pair-wise correlation analysis between the CDSeq
113 estimated GEPs and the scRNA-seq GEPs and assign the CDSeq estimated cell type to the cell
114 type in the scRNA-seq data with which it has the highest correlation. A threshold for the
115 correlation can be set *a priori*. Any CDSeq estimated cell types that failed to achieve the
116 correlation threshold would be considered unannotated. Note that it is possible that more
117 than one CDSeq estimated cell types can match the same cell type.
118
119 Second, cell annotation can also be achieved by clustering both the user-provided scRNA-seq
120 reference data with the pseudo scRNA-seq data generated from the CDSeq estimated cell-type
121 specific gene expression profiles. Under the assumption that the total read count of each CDSeq
122 estimated cell type follows a negative binomial distribution, *cellTypeAssignSCRNA* generates the
123 total synthetic read count for each CDSeq cell type using the CDSeq estimated GEPs for the cell
124 using the R function *rnbinom (n, size, mu)* where *n* can be any integer greater than or equal to 1
125 (See Supplementary Figures S11-S16) and, by default, *mu* (mean) and *size* (dispersion) are
126 estimated from the reference scRNA-seq. These default parameters in *rnbinom* can be
127 replaced by user-provided parameters. Next, *cellTypeAssignSCRNA* generates the synthetic
128 single-cell read counts for each CDSeq cell type using the CDSeq-estimated GEPs and the R
129 function *rmultinom(n, size, prob)*, with *n* being the number of pseudo cells per CDSeq estimated
130 cell type (same as the n in *rnbinom*), *size* set equal to the total read count for the cell type as
131 generated by the preceding call to *rnbinom*, and *prob* provided by CDSeq-estimated cell-type-
132 specific gene expression profile expressed as a multinomial parameter vector (i.e., expression

133 levels expressed as proportions that sum to one). The resulting synthetic scRNA-seq data can

134 then subsequently be combined with user provided scRNA-seq data for clustering analysis and

135 cell type annotation using the Seurat package [25,26]. The clustering algorithm implemented in

136 Seurat is a graph-based approach based on the outcomes of principal components analysis. The

137 number of clusters cannot be set explicitly but is affected by a resolution parameter; the larger

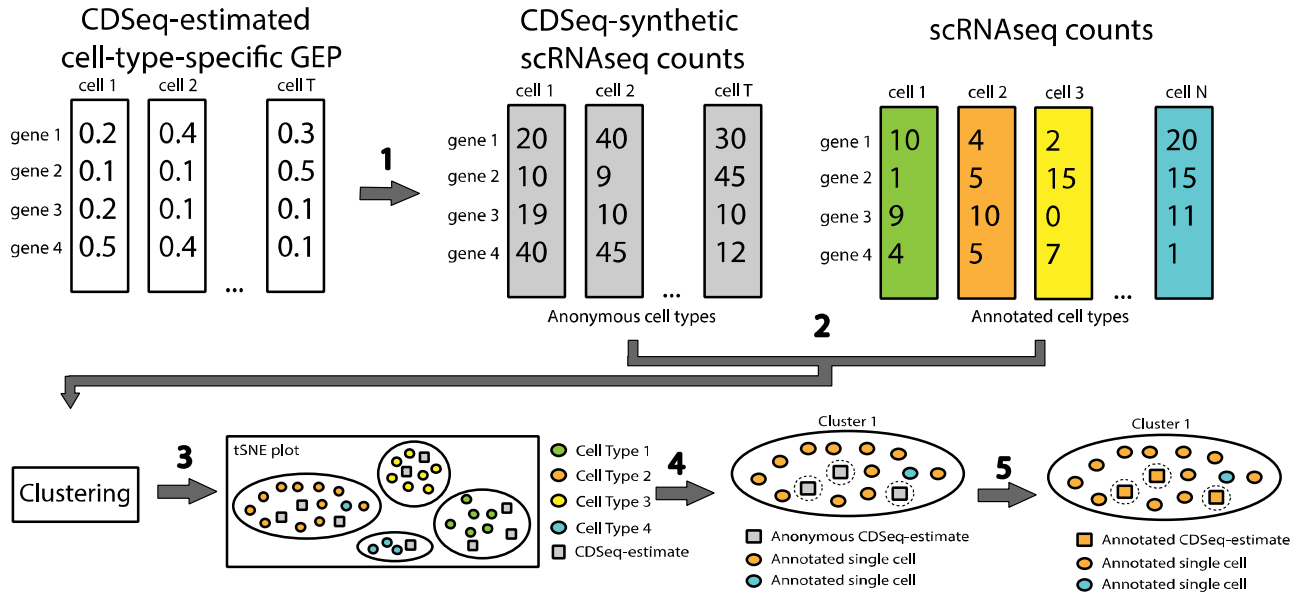
138 the resolution parameter is the more clusters will be returned. The cell types identified by

139 CDSeq from bulk RNA-seq data were clustered with the individual cells from the scRNA-seq data.

140 We then assigned CDSeq-identified cell types to the cell type from the scRNA-seq data to which

141 the majority of the cells in the cluster belong (Fig 1).

142



143

144

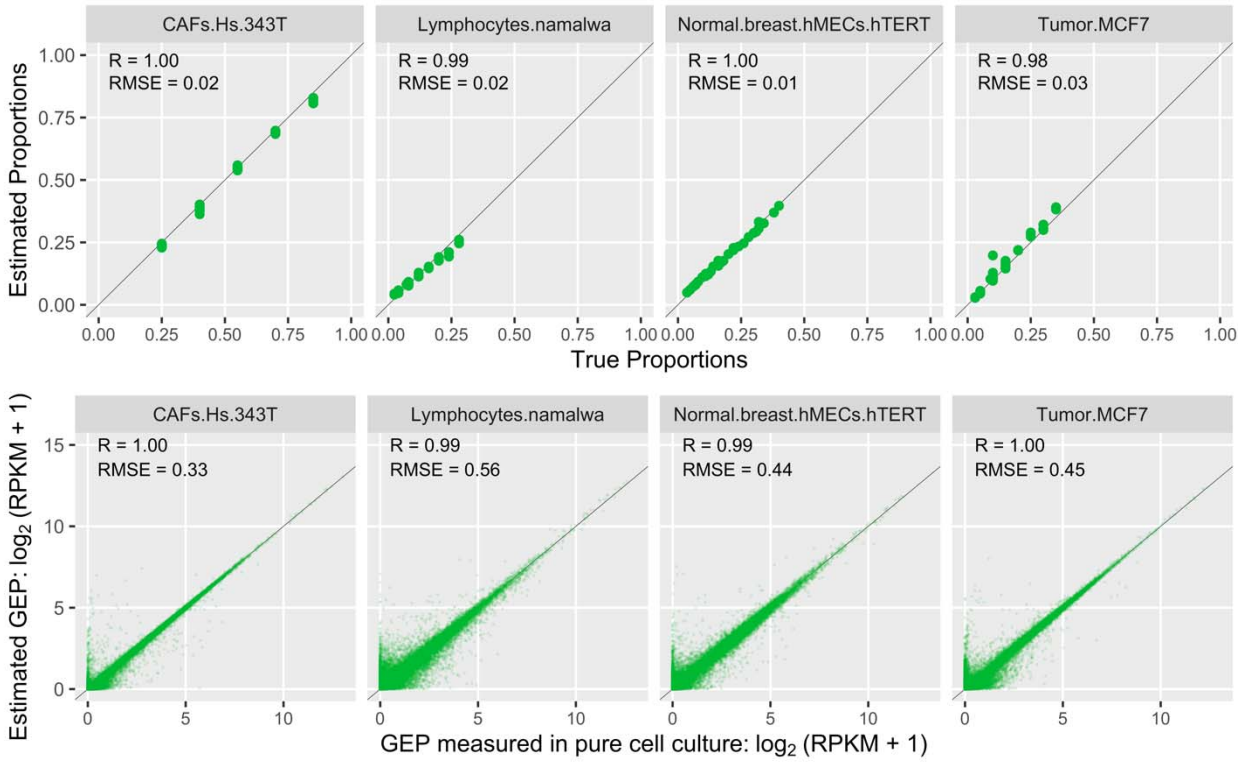
145 Lastly, in the case where the reference data are not available, *cellTypeAssignSCRNA* carried out

146 clustering analysis of the pseudo single-cell data generated from the CDSeq-estimated GEPs

147 using Seurat [25,26] and identifies the marker genes in each cluster for cell type annotation.

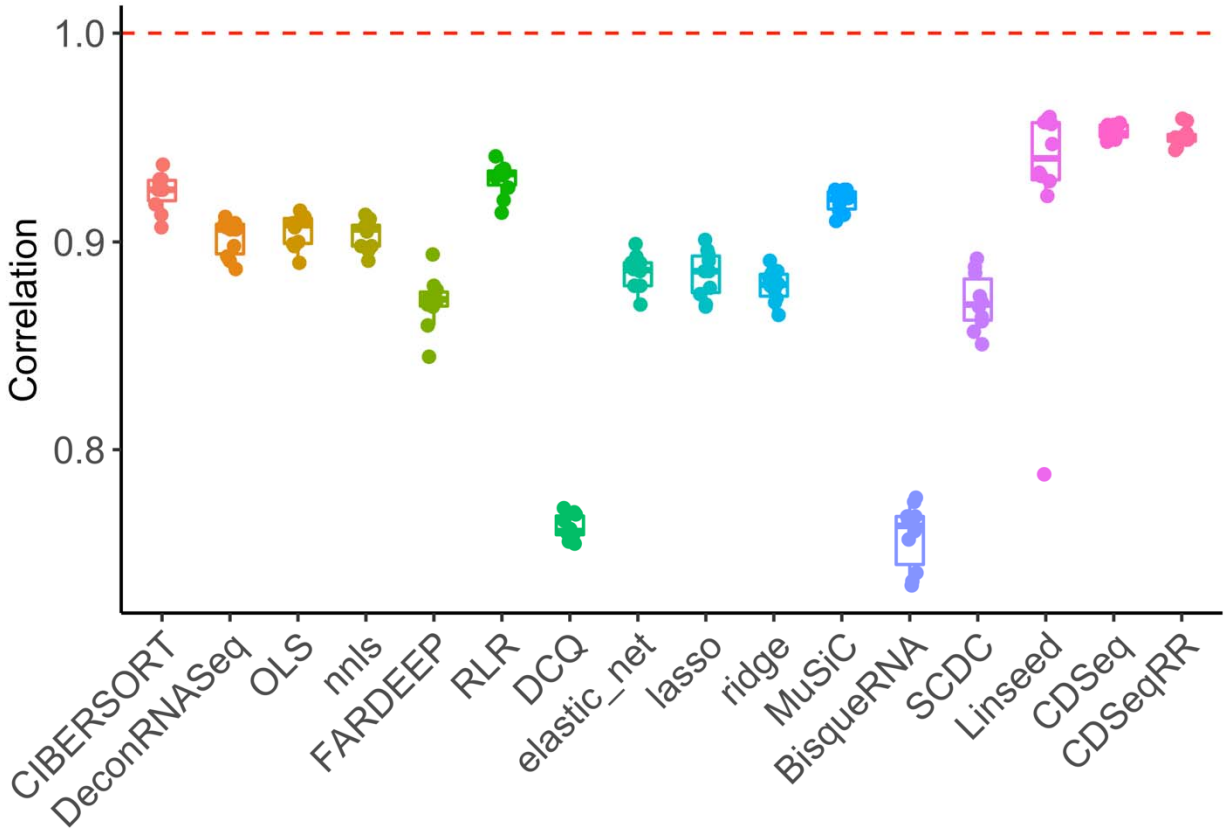
148 The marker genes for each cluster are identified by comparing their expression levels in the
149 cluster with those in all other clusters. We show that this approach works reasonably well (Fig.
150 4 and Supplementary Fig. S17, S18, S19 and S21) using synthetic mixtures of human brain cells
151 [21].
152
153 **Results**
154 We validated CDSeqR using both synthetic and cell mixtures [5] with known cell-type
155 composition and cell-type-specific GEPs. For cell mixtures, the CDSeqR estimates using the
156 Reduce and Recovery strategy matched the known cell-type-specific GEPs and sample-specific
157 cell-type proportions (Fig. 2) while reducing computational time by about 96% (from about 120
158 min to 5 min) compared to the original implementation. In our previous work, we compared
159 CDSeqR with seven other deconvolution methods [5]. Here, we employed a recently published
160 benchmark pipeline [17] to simulate data for extensive comparisons with additional existing
161 deconvolution methods. In total, we compared CDSeqR with 14 competing deconvolution
162 algorithms [6,16,35,27–34] (Fig. 3 and Supplementary Fig. S6-S8) including linseed [6]. CDSeqR
163 software and linseed both provided accurate cell-type proportion using similar amounts of
164 computer time (Supplementary Fig. S1-S5). We further evaluated CDSeq and our new cell type
165 annotation function by examining the results of using different number of cell types in the
166 synthetic mixtures generated from scRNA-seq data (Supplementary Fig. S10-S16). In addition,
167 we examined CDSeqR's ability to detect immune cell subtypes in PBMC mixtures, a challenging
168 problem referred to as deep deconvolution [18,28]. Specifically, we generated synthetic
169 mixtures from 8 immune subtypes: B cells, Cytotoxic T cells, Helper T cells, Memory T cells,

170 Monocytes, Naïve Cytotoxic T cells, Naïve T cells, and Natural killer T cells using scRNA-seq [19]
171 and demonstrated that CDSeqR was able to identify all the immune cell subtypes and provided
172 accurate estimates of cell type proportions (Supplementary Fig. S21 and S22). We further
173 tested CDSeqR using GTEx [36] and TCGA [37] datasets and showed that the CDSeq identified
174 cell types were consistent with the published single-cell RNA-Seq results [20,21]
175 (Supplementary Fig. S23 and S24).



176

177



One input parameter for CDSeq function is the number of cell types to be deconvolved. The choice for the input is data dependent. For example, for a large dataset (e.g., >50), specifying a large number of cell type may help CDSeqR to detect all the major cell types due to the GEP heterogeneity even for a same cell type. Fortunately, CDSeqR allows the user to specify a range of possible cell type numbers as an input vector. For each choice in the vector, CDSeqR will perform computation in parallel and estimate the number of cell types based on the posterior distribution of the underlying model. In all *in silico* mixtures tested, CDSeqR provided accurate estimates for the number of cell types in the synthetic data.

189 In addition, one may use the cell type information in scRNA-seq data from the same tissue type

190 as a guide. Here we showed that using synthetic mixture datasets generated from scRNA-seq

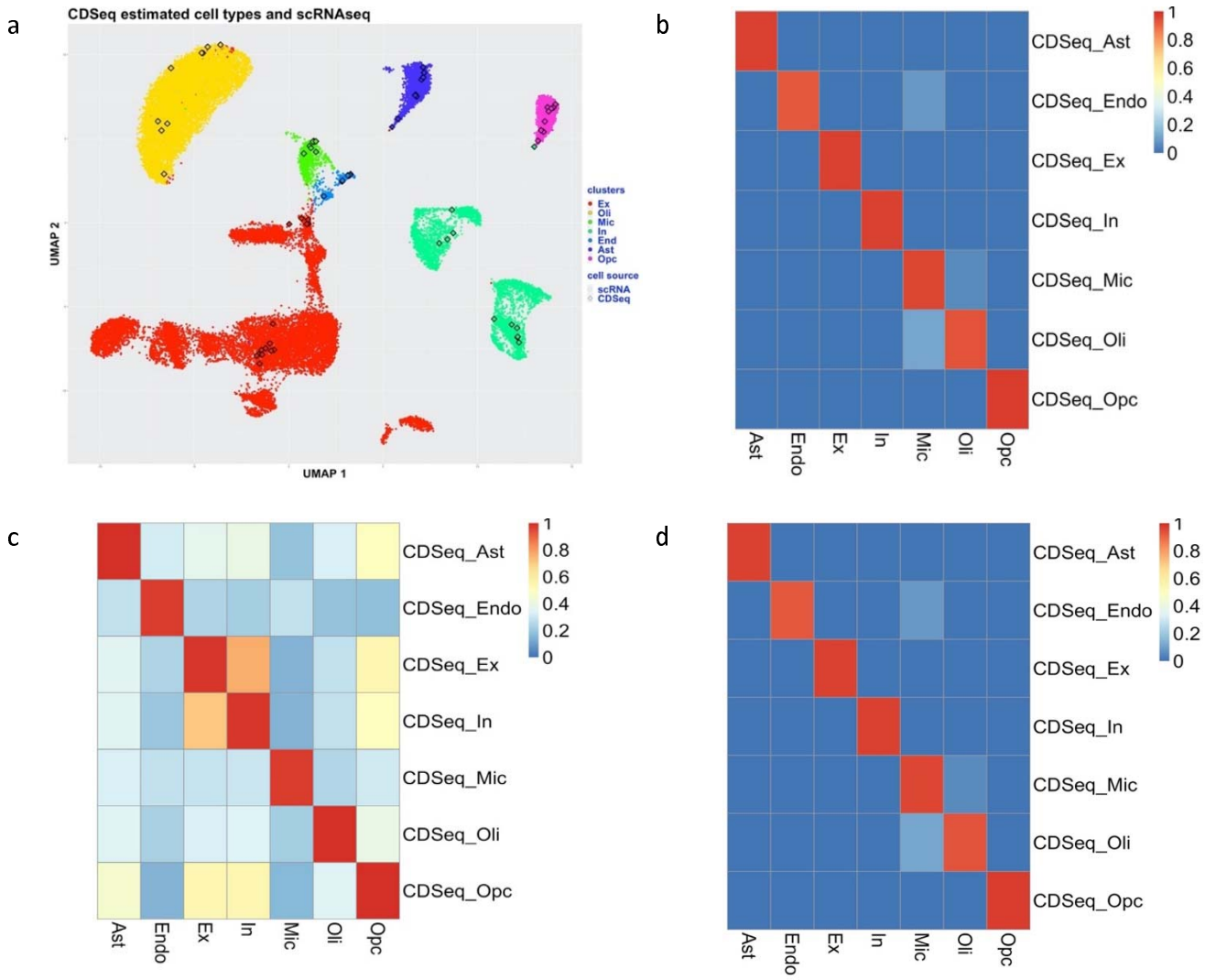
191 data, the newly implemented cell type annotation function can successfully uncover the cell

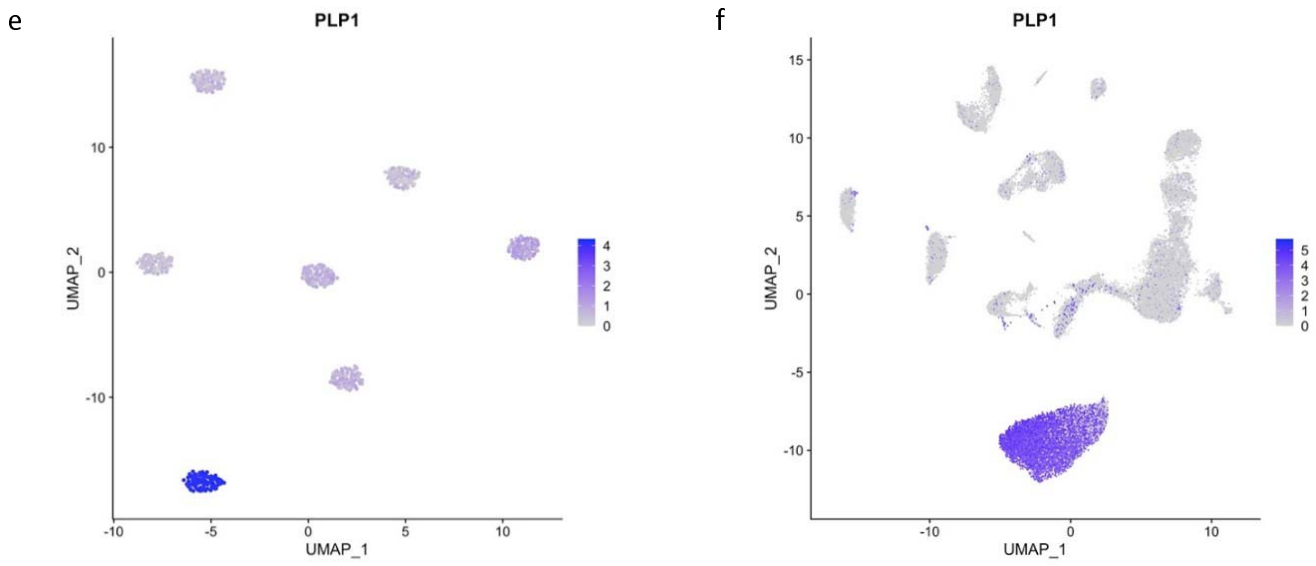
192 types (Fig. 4 and Supplementary fig. S10-24). Our simulation studies suggested that when the

193 specified cell type number is larger than the true value, the cell type annotation function is able

194 to identify and combine those cell types that are similar into one (Supplementary Fig. 10-16).

195





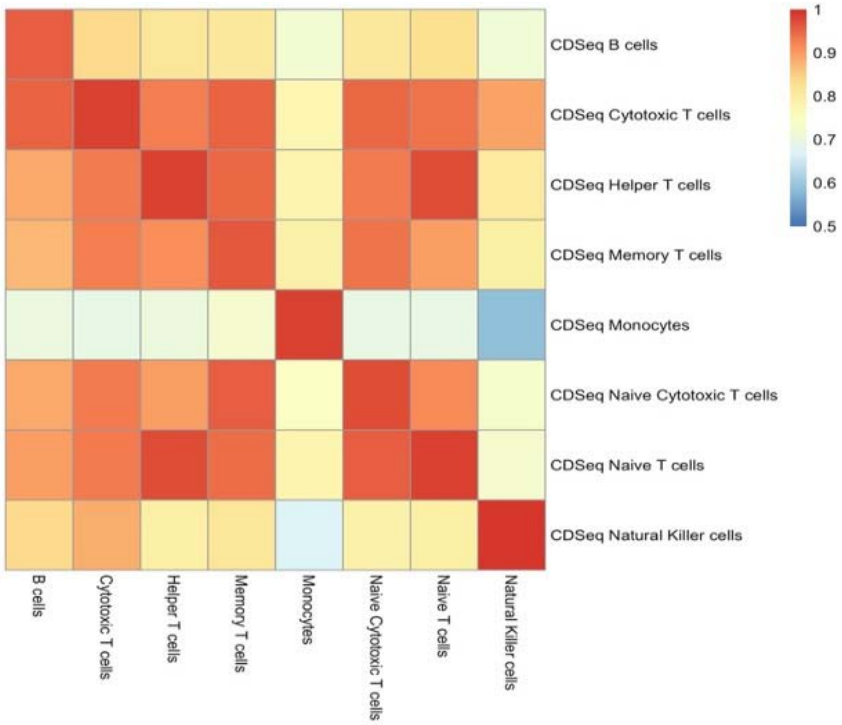
196

197 Another challenge for deconvolution is deep deconvolution [28], namely, to infer component
198 subtypes from bulk tissues. For example, researchers are often interested in uncovering
199 immune cell subtypes among major immune cell populations. In our previous work [18], we
200 demonstrated that CDSeq was able to estimate immune cell subtypes from PBMC (peripheral
201 blood mononuclear cell) data using microarray data. In this work, we further investigated
202 CDSeq's ability to infer immune cell subtypes using synthetic mixtures generated from scRNA-
203 seq data [19]. Using a deconvolution benchmarking pipeline [17], we created 100 synthetic
204 mixtures of various proportions of the 8 immune cell subtypes: B cells, Monocytes, Helper T
205 cells, Natural killer cells, Cytotoxic T cells, Memory T cells, Naïve Cytotoxic T cells, and Naïve T
206 cells. The highly-correlated GEPs of several of the cell types present major challenge for
207 deconvolution methods. To test and validate the performance of CDSeqR, we ran CDSeqR by
208 specifying a range of varying numbers of cell types from 2 to 40. CDSeqR can automatically
209 identify similar cell types and merge them into one. CDSeqR was able to identify all 8 immune

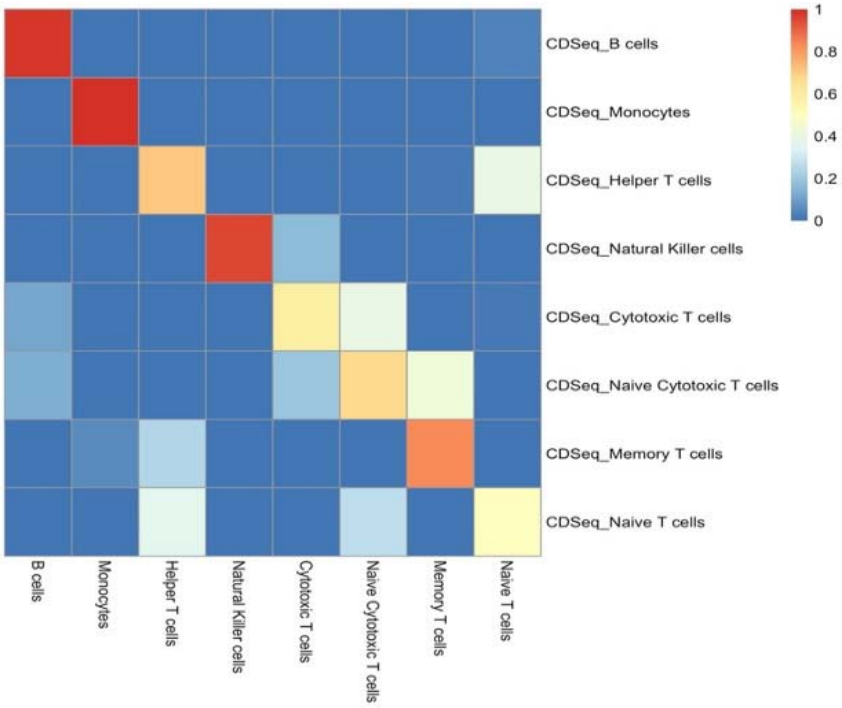
210 cell subtypes when the input number of the cell types was set to be greater than 10 (Fig 5;

211 Supplementary Fig. S21 and S22).

a



b



212

213 Besides the synthetic data, to further showcase the utility of CDSeqR and the newly
214 implemented cell annotation function, we carried out deconvolution analysis of the RNA-seq
215 data from human frontal cortex (209 samples) from the GTEx V8 [36]. We used publicly
216 available brain single-cell RNA-seq data [21] as the reference for cell annotation. We set the
217 number of cell types from 50 to 100 and ran CDSeqR with a dilution factor of 100 to facilitate
218 computation. The result demonstrates that setting the number of cell types to 100 enabled
219 detection of the major cell types in heterogenous brain samples whereas a smaller number (e.g.,
220 50) failed to identify oligodendrocyte progenitor cell, a known major cell type (Supplementary
221 Fig.S23). Setting the cell types to be 100 is reasonable as the brain is heterogenous and cell
222 states are known to represent a continuum. This notion is consistent with the finding that, in
223 brain prefrontal cortex [21,38], many of the major cell types consist of multiple subtypes. For
224 example, excitatory neurons have more than ten subtypes [21,38]. In a second example, we
225 deconvolved TCGA tumor samples without using any references. See Supplementary Fig.S24 for
226 details.
227
228 **Discussion**
229 CDSeq is a complete deconvolution method that uses gene expression data alone to estimate
230 both sample-specific cell-type proportions and cell-type-specific gene expression profiles from
231 RNA-seq data [5]. CDSeq was originally developed in MATLAB. In this work, we developed an R
232 package, CDSeqR, for a broader group of users. Moreover, in the R implementation, we
233 substantially improved CDSeq's computational efficiency in both speed and memory usage.
234 Specifically, we proposed a novel strategy – Reduce and Recover – to allow CDSeq to run on

235 multiple processors simultaneously (parallelization). We further validated our CDSeqR package
236 using data from both *in silico* and *in vitro* cell mixtures. We compared the performance of
237 CDSeqR with that of 14 competing deconvolution method and showed that cell-type
238 proportions estimated by CDSeqR are as or more accurate than those estimated by the
239 competing methods. However, unlike many competing methods CDSeqR does not require
240 reference GEPs for deconvolution and also provides cell-type-specific GEP estimates. We
241 provided recommendations for parameter settings in Supplementary Table S1 and the
242 parameters we used in our results in Supplementary Table S2.
243
244 In the R package, we also implemented a new cell-type assignment function. This new function
245 takes advantage of the publicly available scRNA-seq datasets and use them as a reference to aid
246 the annotation and visualization of the CDSeq estimated cell types. Specifically, CDSeqR
247 transforms its estimated cell-type-specific GEPs into pseudo single-cell read counts (see
248 Methods). Next, the two single-cell datasets (i.e. the external experimental scRNA-seq and the
249 pseudo-scRNA-seq generated from CDSeq estimates) are combined and clustered using existing
250 software such as Seurat [25,26]. Lastly, CDSeqR annotates the CDSeq cell types using either
251 clustering analysis or correlation analysis (See Methods). In the case where no reference scRNA-
252 seq dataset is available, CDSeqR identifies marker genes and uses them to aid cell type
253 annotation.
254
255 The ongoing discovery of new cell subtypes [21,23,39] suggest that cell types may be better
256 described as a continuum rather than as a sequence of discrete states. This recognition,

consequently, brings challenges to deconvolution algorithms, especially for those that rely reference gene expression profiles [4,16,28,40]. The references will not fully capture the heterogeneity and the continuum of states exhibited by cell types in bulk tissues. Compared to existing deconvolution methods, CDSeq employs an unsupervised framework (does not require a reference panel) and can discover cell types in the data. Nonetheless, CDSeqR can only predict discrete cell types, not cell continuum. However, we showed that CDSeqR can do more than detect only major cell types but also can help users identify immune subtypes.

Since cell types represent a continuum, it is difficult to assign a value to the number of cell types present. This issue is analogous to determining the number of clusters from a clustering analysis. Nonetheless, we suggest that, in practice, one may set the number of cell types to be a vector that covers a range of plausible values. For example, set parameter cell_type_number = seq(20,100,10)) will allow CDSeq to try the number of cell types from 20 to 100 with an increment of 10. One can then visualize the clustering results with experimental scRNA-seq data using the cell-type annotation function and also consider the values of log posterior reported by CDSeq as a reference to determine the appropriate number of cell types. Importantly, the underlying biology would need to be considered when determining the number cell types in the data.

Conclusion

We developed an R package for our complete deconvolution method [5] and implemented new functions to improve computational efficiency. Our method employs an unsupervised learning

279 framework and requires no reference profile for the deconvolution procedure. We provided
280 multiple ways to help identify CDSeq-estimated cell types after deconvolution. When a reliable
281 single cell reference dataset is available, one can annotate deconvolved cell types using
282 correlation analysis or clustering analysis. In the case where a single cell reference is not
283 available, CDSeqR will identify marker genes for each deconvolved cell type to help users
284 annotate the identities of those cell types. Our method is also able to infer the number of cell
285 types from bulk RNA-seq measurements. In summary, we believe that our CDSeqR package will
286 be a valuable tool for computationally deciphering sample heterogeneity using bulk RNA-Seq
287 data.
288
289 **Abbreviations**
290 TCGA: The Cancer Genome Atlas
291 GTEx: The Genotype-Tissue Expression (GTEx) project
292
293 **Availability and requirements**
294 Project name: CDSeq R package
295 Project home page: https://github.com/kkang7/CDSeq_R_Package
296 Operating system: Platform independent
297 Programming language: R, C++, C.
298 Other requirements: None
299 License: GPL-3
300 Any restrictions to use by non-academics: None

Declarations

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Availability of data and materials

Source code are available on Github (https://github.com/kkang7/CDSeq_R_Package) and CRAN (https://cran.r-project.org/web/packages/CDSeq/index.html).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences (ES101765). The funding body plays no roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

LL, KK and DU conceived the project. KK designed the new strategy for parallelization and developed the new function for annotating CDSeq-estimated cell types. KK and DH implemented the R package. KK, DH and YL performed numerical testing and validations. All authors contributed to the writing of the manuscript. LL guided the research. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the NIEHS Office of Scientific Computing for computing support and the NIEHS Epigenomics and DNA Sequencing Core Facility for RNA-seq analysis of the cell mixture samples.

Reference

1. Demaria O, Cornen S, Daëron M, Morel Y, Medzhitov R, Vivier E. Harnessing innate immunity in cancer therapy. Nature. 2019.
2. Zheng C, Zheng L, Yoo JK, Guo H, Zhang Y, Guo X, et al. Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. Cell. 2017;
3. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, et al. Cell type--specific gene expression differences in complex tissues. Nat Methods. 2010;7(4):287–9.
4. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;
5. Kang K, Przytycka TM, Meng Q, Shats I, Umbach DM, Li M, et al. CDSeq: A novel complete

345 deconvolution method for dissecting heterogeneous samples using gene expression data.
346 PLOS Comput Biol [Internet]. 2019 Dec 2; Available from:
347 https://doi.org/10.1371/journal.pcbi.1007510
348 6. Zaitsev K, Bambouskova M, Swain A, Artyomov MN. Complete deconvolution of cellular
349 mixtures based on linearity of transcriptional signatures. Nat Commun. 2019;
350 7. Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific
351 information from heterogeneous samples. Curr Opin Immunol. 2013;25(5):571–8.
352 8. Erkkilä T, Lehmusvaara S, Ruusuvuori P, Visakorpi T, Shmulevich I, Lähdesmäki H.
353 Probabilistic analysis of gene expression measurements from heterogeneous tissues.
354 Bioinformatics. 2010;
355 9. Qiao W, Quon G, Csaszar E, Yu M, Morris Q, Zandstra PW. PERT: a method for expression
356 deconvolution of human blood samples from varied microenvironmental and
357 developmental conditions. PLoS Comput Biol. 2012;8(12):e1002838.
358 10. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of
359 heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics.
360 2013;29(8):1083–5.
361 11. Zhong Y, Wan YW, Pang K, Chow LML, Liu Z. Digital sorting of complex tissues for cell
362 type-specific gene expression profiles. BMC Bioinformatics. 2013;
363 12. Gaujoux R, Seoighe C. Semi-supervised Nonnegative Matrix Factorization for gene
364 expression deconvolution: a case study. Infect Genet Evol. 2012;12(5):913–21.
365 13. Li Y, Xie X. A mixture model for expression deconvolution from RNA-seq in
366 heterogeneous tissues. BMC Bioinformatics. 2013;14(5):S11.

367 14. Ahn J, Yuan Y, Parmigiani G, Suraokar MB, Diao L, Wistuba II, et al. DeMix: Deconvolution
368 for mixed cancer transcriptomes using raw measured data. Bioinformatics. 2013;
369 15. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of
370 cell subsets from tissue expression profiles. Nat Methods. 2015;
371 16. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-
372 subject single-cell expression reference. Nat Commun. 2019;
373 17. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking
374 of cell type deconvolution pipelines for transcriptomics data. Nat Commun. 2020;
375 18. Kang K, Meng Q, Shats I, Umbach DM, Li M, Li Y, et al. CDSeq: A novel complete
376 deconvolution method for dissecting heterogeneous samples using gene expression data.
377 PLoS Comput Biol. 2019;
378 19. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel
379 digital transcriptional profiling of single cells. Nat Commun. 2017;
380 20. Puram S V., Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell
381 Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck
382 Cancer. Cell. 2017;
383 21. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell
384 transcriptomic analysis of Alzheimer's disease. Nature. 2019;
385 22. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci. 2004;
386 23. Pastushenko I, Brisebarre A, Sifrim A, Fioramonti M, Revenco T, Boumahdi S, et al.
387 Identification of the tumour transition states occurring during EMT. Nature. 2018;
388 24. Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, et al. Construction of a human cell landscape

389 at single-cell level. Nature. 2020;
390 25. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic
391 data across different conditions, technologies, and species. Nat Biotechnol. 2018;
392 26. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al.
393 Comprehensive Integration of Single-Cell Data. Cell. 2019;
394 27. Chambers JM, Hastie TJ. Statistical models in S. Statistical Models in S. 2017.
395 28. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of
396 cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453.
397 29. Du R, Carey V, Weiss ST. DeconvSeq: Deconvolution of cell mixture distribution in
398 sequencing data. Bioinformatics. 2019;
399 30. Hao Y, Yan M, Heath BR, Lei YL, Xie Y. Fast and robust deconvolution of tumor infiltrating
400 lymphocyte from expression profiles using least trimmed squares. PLoS Comput Biol.
401 2019;
402 31. Riplley B, Venables B, Bates DM, Firth D, Hornik K, Gebhardt A. Package "MASS". Support
403 Functions and Datasets for Venables and Ripley's MASS. Document freely available on
404 the internet at: http://www. r-project. org. 2018.
405 32. Altboum Z, Steuerman Y, David E, Barnett-Itzhaki Z, Valadarsky L, Keren-Shaul H, et al.
406 Digital cell quantification identifies global immune cell dynamics during influenza
407 infection. Mol Syst Biol. 2014;
408 33. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via
409 coordinate descent. J Stat Softw. 2010;
410 34. Jew B, Alvarez M, Rahmani E, Miao Z, Ko A, Garske KM, et al. Accurate estimation of cell

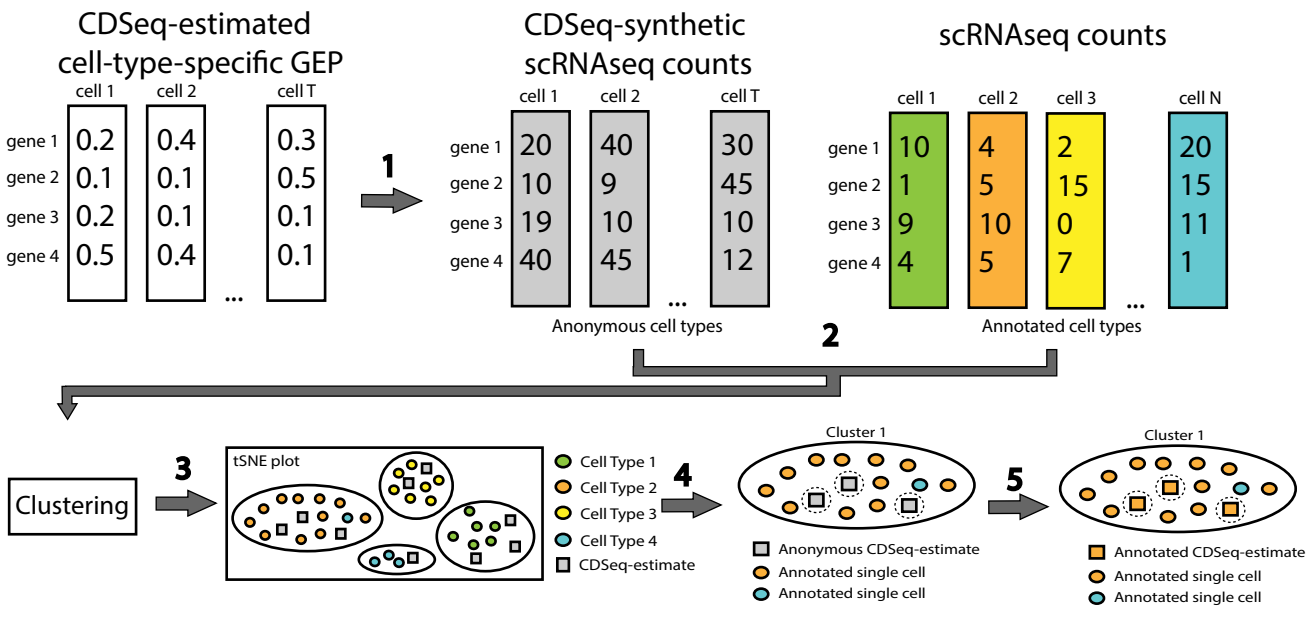
411 composition in bulk expression through robust integration of single-cell information. Nat
412 Commun. 2020;
413 35. Dong M, Thennavan A, Urrutia E, Li Y, Perou CM, Zou F, et al. SCDC: bulk gene expression
414 deconvolution by multiple single-cell RNA sequencing references. Brief Bioinform. 2021;
415 36. Aguet F, Barbeira AN, Bonazzola R, Brown A, Castel SE, Jo B, et al. The GTEx Consortium
416 atlas of genetic regulatory effects across human tissues. Science (80-). 2020;
417 37. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The
418 cancer genome atlas pan-cancer analysis project. Nature Genetics. 2013.
419 38. Gandal MJ, Zhang P, Hadjimichael E, Walker RL, Chen C, Liu S, et al. Transcriptome-wide
420 isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science (80-).
421 2018;
422 39. Trapnell C. Defining cell types and states with single-cell genomics. Genome Res.
423 2015;25(10):1491–8.
424 40. Hunt GJ, Freytag S, Bahlo M, Gagnon-Bartsch JA. Dtangle: Accurate and robust cell type
425 deconvolution. Bioinformatics. 2019;
426 41. Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, et al.
427 Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat
428 Biotechnol. 2020;
429 42. McKenzie AT, Wang M, Hauberg ME, Fullard JF, Kozlenkov A, Keenan A, et al. Brain Cell
430 Type Specific Gene Expression and Co-expression Network Architectures. Sci Rep. 2018;
431

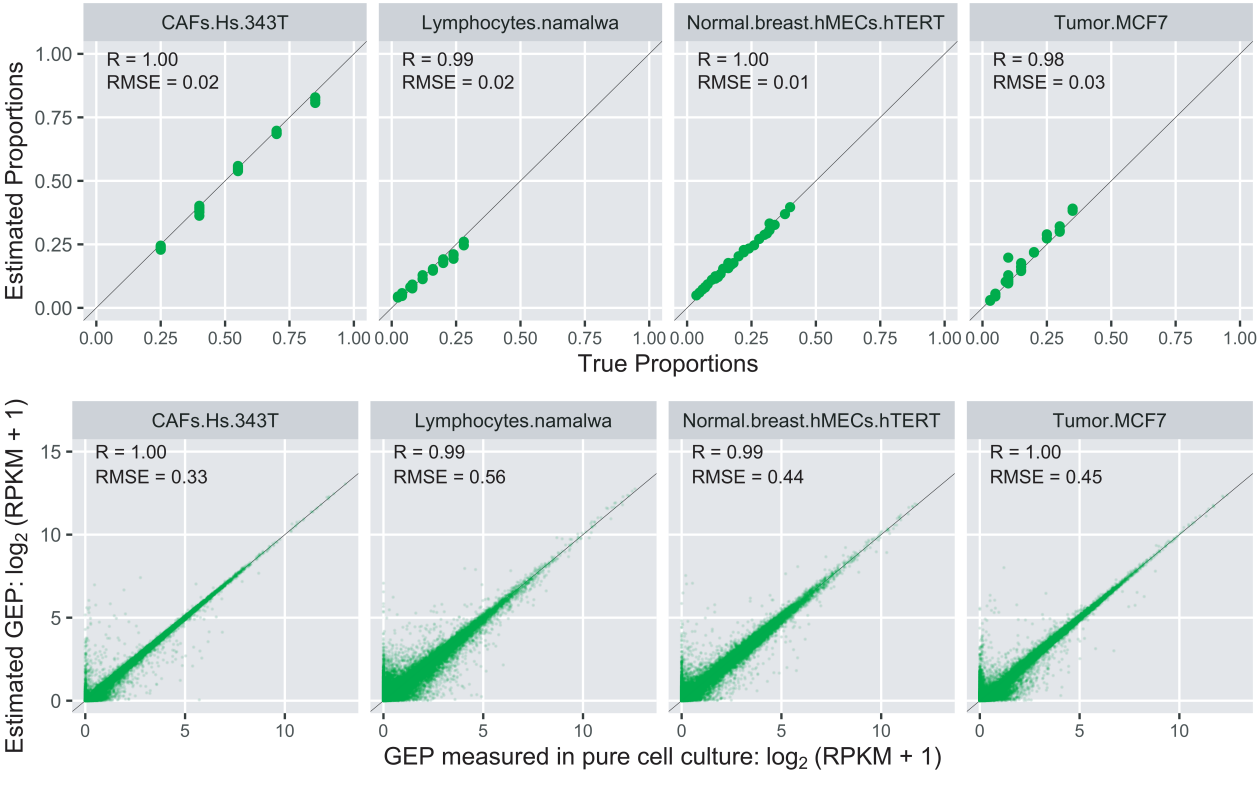
432 **Figure legend**

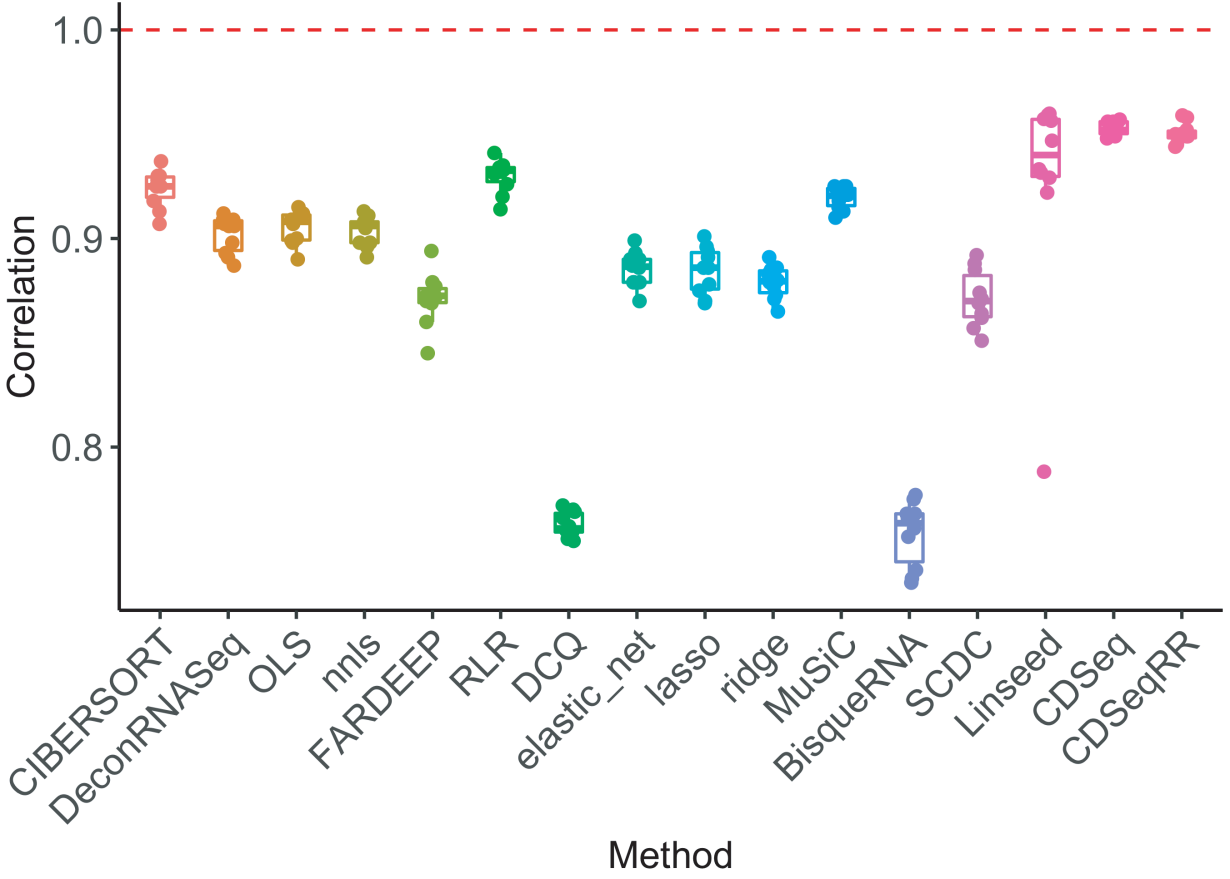
433 Figure 1. Schematic illustration of our cell-type assignment approach. It consists of five steps: 1)
434 generating synthetic single-cell read counts (gray rectangles) for the CDSeq-estimated cell-type-
435 specific GEPs (white rectangles); 2) combining the pseudo scRNA-seq data with user-provided
436 scRNA-seq data (colored rectangles), then performing clustering using Seurat; 3) visualizing the
437 result using tSNE plot to show the 2D embedding of the clustering outcomes with differently
438 colored circles denoting distinctive cell types in the single-cell data and gray squares denoting
439 the CDSeq-identified cell types; 4) annotating each cluster to the cell type to which the majority
440 of the cells in the cluster belong; 5) assigning the CDSeq-identified cell types accordingly as in
441 step 4.
442
443 Figure 2. Results of CDSeqR analysis of the *in vitro* cell mixtures. The mixtures comprise 4 four
444 cell types: cancer-associated fibroblast, lymphocytes, normal breast cells, tumor cells (see
445 Supplementary Methods for details). We ran CDSeqR on 32 bulk RNA-seq samples with known
446 cell proportions in each sample [18] using 7 randomly sampled gene subsets of size 500 (7 data
447 blocks, each of 500 genes). In the first row, each point represents a sample (n=32); in the
448 second row, each point represents a gene (n=~19,000). True and estimated GEPs are
449 normalized as RPKM. The line in each panel is the x=y line.
450
451 Figure 3. Comparisons with existing deconvolution methods using data generated from
452 benchmark pipeline [17]. We compared CDSeqR with 14 other deconvolution methods using
453 synthetic mixtures generated from three scRNA-seq datasets [20,21,41]. The results for the
454 PBMC dataset [41] are shown here. The X-axis represents different deconvolution methods

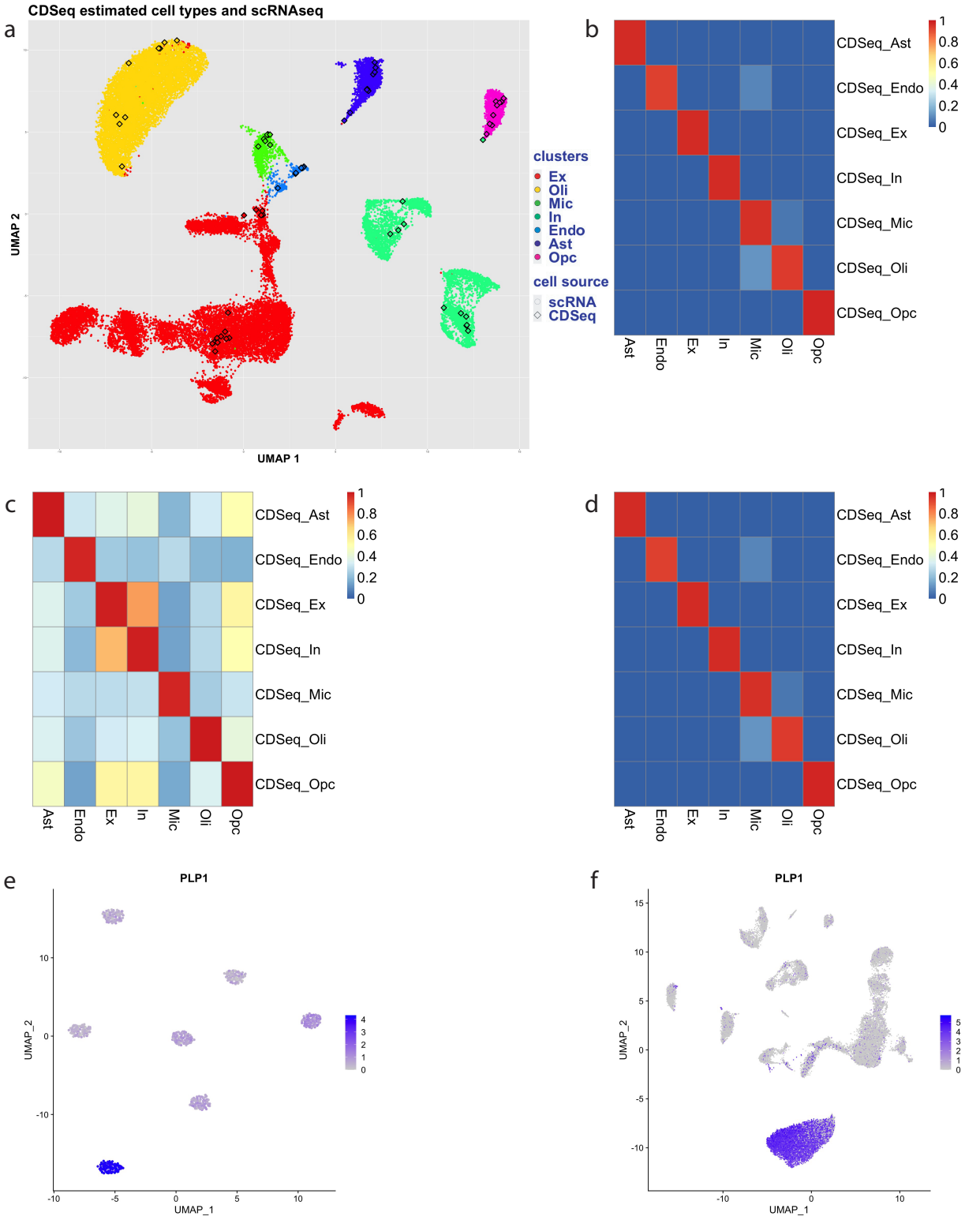
455 whereas the Y-axis denotes the correlation between the estimated cell type proportions by
456 each method and the true cell type proportions. We generated 10 sets of 100 mixtures. Each
457 dot in the box plots represents a deconvolution result from one of the 10 sets. CDSeqRR refers
458 to deconvolution with the Reduce-Recover strategy whereas CDSeq denotes deconvolution
459 without using Reduce-Recover option Additional results of the comparisons are available in
460 Supplementary Fig. S6-S8.
461
462 Figure 4. Annotation of CDSeq-identified cell types. CDSeqR was run with the Reduce-Recover
463 option using 800 genes per block and 10 blocks. A reference scRNA-seq dataset was provided
464 for cell type annotation using clustering analysis (a-b) and correlation analysis (c-d). (a) UMAP
465 embeddings of reference scRNA-seq data (color dots) and pseudo single-cell data generated
466 from the CDSeq-estimated cell-type specific gene expression profiles (color squares). Each
467 distinct color in the UMAP plot denotes a cell type provided with the scRNA-seq annotation
468 given in [21]. There are 7 major cell types provided with the scRNA-seq: excitatory neuron (Ex),
469 inhibitory neuron (In), oligodendrocyte (Oli), microglia (Mic), endothelial (Endo), astrocyte (Ast)
470 and oligodendrocyte progenitor cell (OPC). The deconvolution result used 7 as the number of
471 cell types and for each CDSeq-estimated cell type, it generated 10 pseudo cells which
472 correspond to a color square in the UMAP plot. The colors of the squares (i.e. CDSeq estimated
473 cell types) were determined based on its annotation result. (b) correlations between the
474 resulting CDSeqR estimated cell type proportions and true proportions; (c) correlations
475 between the CDSeqR estimated cell-type-specific GEPs and scRNA-seq GEPs; (d) correlation of
476 CDSeq-estimated cell type proportions with true proportions using correlation analysis (see

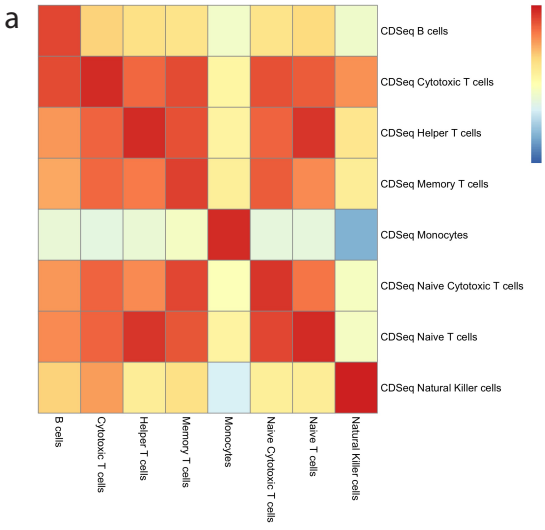
477 Methods for details). When reference for cell annotation is not provided, CDSeq will perform
478 clustering analysis on the pseudo single-cell data generated from the CDSeq-estimated cell-type
479 specific gene expression profiles and perform differential gene expression analysis to identify
480 the marker genes in each cluster using the Seurat software [25,26]. (e) Expression level of PLP1
481 (Oligodendrocyte marker [42]) in the clusters of CDSeq-estimated cell types (cell type number is
482 7); (f) Expression level of PLP1 in human brain prefrontal cortex scRNA-seq clusters [21]. See
483 Supplementary Figure S10-S20 for more details.
484
485 Figure 5. Results of deconvolution of PBMC immune cell subtype mixtures using Reduce-
486 Recover strategy. The number of cell types was set to be 20. (a) Correlation between the CDSeq
487 estimated cell-type-specific GEPs and true cell-type-specific GEPs; (b) Correlation between
488 CDSeq estimated sample-specific cell-type proportions and true cell type proportions.
489 Additional results for different specifications of the cell types are provided in Supplementary Fig.
490 S21 and S22.
491
492 **Additional files**
493 File name: Additional file 1
494 Title: Supplementary Note for CDSeqR
495 Description: Supplementary methods, tables and figures

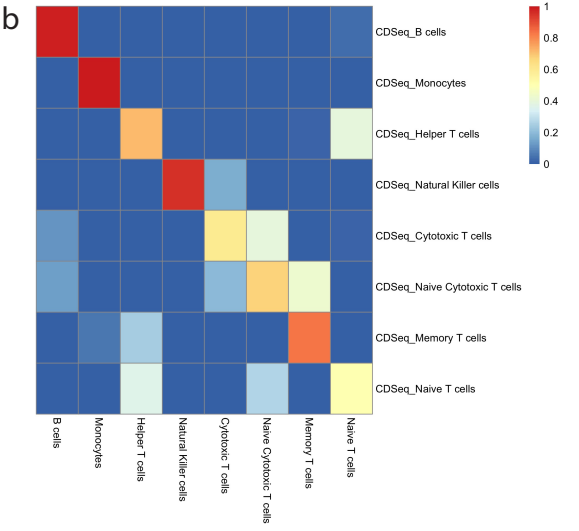

CDSeq-estimated
cell-type-specific GEP
cell 1
cell 2
cell T
gene 1
gene 2
gene 3
gene 4
0.2
0.1
0.2
0.5
0.4
0.1
0.1
0.4
0.3
0.5
0.1
0.1
...
1
CDSeq-synthetic
scRNAseq counts
cell 1
cell 2
cell T
gene 1
gene 2
gene 3
gene 4
20
10
19
40
40
9
10
45
30
45
10
12
...
Anonymous cell types
scRNAseq counts
cell 1
cell 2
cell 3
cell N
gene 1
gene 2
gene 3
gene 4
10
1
9
4
4
5
10
5
2
15
0
7
20
15
11
1
...
Annotated cell types
2
Clustering
3
tSNE plot
Cell Type 1
Cell Type 2
Cell Type 3
Cell Type 4
CDSeq-estimate
4
Cluster 1
Anonymous CDSeq-estimate
Annotated single cell
Annotated single cell
5
Cluster 1
Annotated CDSeq-estimate
Annotated single cell
Annotated single cell


CAFs.Hs.343T
R = 1.00
RMSE = 0.02
Lymphocytes.namalwa
R = 0.99
RMSE = 0.02
Normal.breast.hMECs.hTERT
R = 1.00
RMSE = 0.01
Tumor.MCF7
R = 0.98
RMSE = 0.03
Estimated Proportions
True Proportions
0.00
0.25
0.50
0.75
1.00
CAFs.Hs.343T
R = 1.00
RMSE = 0.33
Lymphocytes.namalwa
R = 0.99
RMSE = 0.56
Normal.breast.hMECs.hTERT
R = 0.99
RMSE = 0.44
Tumor.MCF7
R = 1.00
RMSE = 0.45
Estimated GEP: log2 (RPKM + 1)
GEP measured in pure cell culture: log2 (RPKM + 1)
0
5
10
15


1.0
0.9
0.8
Correlation
CIBERSORT
DeconRNASeq
OLS
nnls
FARDEEP
RLR
DCQ
elastic_net
lasso
ridge
MuSiC
BisqueRNA
SCDC
Linseed
CDSeq
CDSeqRR
Method


CDSeq estimated cell types and scRNAseq
a
UMAP 2
UMAP 1
clusters
Ex
Oli
Mic
In
Endo
Ast
Opc
cell source
scRNA
CDSeq
b
CDSeq_Ast
CDSeq_Endo
CDSeq_Ex
CDSeq_In
CDSeq_Mic
CDSeq_Oli
CDSeq_Opc
Ast
Endo
Ex
In
Mic
Oli
Opc
1
0.8
0.6
0.4
0.2
0
c
d
e
PLP1
UMAP_2
UMAP_1
f


a
CDSeq B cells
CDSeq Cytotoxic T cells
CDSeq Helper T cells
CDSeq Memory T cells
CDSeq Monocytes
CDSeq Naive Cytotoxic T cells
CDSeq Naive T cells
CDSeq Natural Killer cells
B cells
Cytotoxic T cells
Helper T cells
Memory T cells
Monocytes
Naive Cytotoxic T cells
Naive T cells
Natural Killer cells
1
0.9
0.8
0.7
0.6
0.5


b
CDSeq_B cells
CDSeq_Monocytes
CDSeq_Helper T cells
CDSeq_Natural Killer cells
CDSeq_Cytotoxic T cells
CDSeq_Naive Cytotoxic T cells
CDSeq_Memory T cells
CDSeq_Naive T cells
B cells
Monocytes
Helper T cells
Natural Killer cells
Cytotoxic T cells
Naive Cytotoxic T cells
Memory T cells
Naive T cells
1
0.8
0.6
0.4
0.2
0