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, Durham, North Carolina, 27709, USA

[*]To whom correspondence should be addressed.

[†]These authors contribute equally.

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 with a new function to aid interpretation of deconvolution outcomes. 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 CDSeq-estimated cell types using publicly available single-cell RNA sequencing (scRNA-seq) data (single-cell data from 20 major organs are included in the R package). This function allows users to readily interpret and visualize the CDSeq-estimated cell types. We carried out additional validations of the CDSeqR software with *in silico* and *in vitro* mixtures and with real experimental data including 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. However, bulk level analyses neglect tissue heterogeneity and hinder investigation in a cell-type-specific fashion. The CDSeqR package can be viewed as providing *in silico* single-cell dissection of bulk measurements. It enables researchers to gain cell-type-specific information from bulk RNA-seq data.

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

45 study how individual cell types in the tissue microenvironment in influence the underlying
46 biology of interest. Computational deconvolution methods [3–16] are complementary to single-
47 cell analysis and are very useful tools for extracting cell-type-specific signals from bulk
48 measurements.
49
50 This work describes an R package (CDSeqR) implementation of CDSeq (complete deconvolution
51 using sequencing data) that we recently developed in MATLAB. CDSeq [9] was designed for
52 simultaneously estimating cell-type-specific gene expression profiles (GEPs) and sample-specific
53 cell-type proportions using bulk RNA-seq data without providing reference GEPs (i.e. pure cell
54 line expressions or scRNA-seq data) to the deconvolution procedure. Here, we not only
55 implemented a user-friendly R version for CDSeq but also substantially improved its
56 computational efficiency in both speed and memory usage. Specially, we designed a dimension
57 Reduction-Recover approach that is automatically parallelized to facilitate computation. In
58 addition, we implemented a new function for assigning the CDSeq-estimated cell types using
59 publicly available scRNA-seq data. Furthermore, we thoroughly tested CDSeqR using both
60 simulated and real benchmark data [9].
61
62 **Implementation**
63 **Statistical model**
64 CDSeq employs a hierarchical Bayesian modeling strategy to account for the mixing process of
65 observed RNA reads from different cell types. Specifically, the gene expression of any particular
66 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 [9,17] for
69 parameter estimation. A detailed description of the model and solver can be found in [9].
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-
75 dependent; therefore, it can be computationally demanding especially for large datasets. We
76 propose a novel strategy for alleviating this issue. Our strategy, named dimension Reduce-
77 Recover, consists of two steps: dimension reduction and dimension recovery. This approach
78 was based on the following assumption: cell proportions in each sample are invariant with
79 respect to the genes used in deconvolution. Specifically, for a bulk RNA-seq sample, the
80 proportions of cell types can be estimated using multiple subsets of randomly selected genes.
81 Therefore, for a bulk RNA-Seq dataset, we should be able to obtain the estimated cell-type
82 proportions for each sample in the dataset regardless of the subset of genes considered. This
83 idea allows us to perform multiple independent estimations of cell-type proportions using
84 many subsets of randomly selected genes processed in parallel. The final estimate of the cell-
85 type proportions is the average of the independent estimates. We then estimate the cell-type
86 specific gene expression profiles by post-multiplying the input data matrix by a generalized
87 inverse of the matrix of final estimated cell-type proportions (Supplementary Methods). We

88 have carried out thorough testing and demonstrated that this approach substantially speeds up
89 computation without sacrificing performance (see Supplementary Note for details).
90
91 **A new cell type assignment function in CDSeqR**
92 Deconvolution results from bulky tissue RNA-seq samples usually present a greater degree of
93 heterogeneity than those from cell mixtures or *in silico* experiments. We have found that, for
94 comprehensive cell-type discovery, the algorithm's parameter governing number of cell types
95 to be deconvolved must be set large. For example, cancer cells in a tumor exhibit a spectrum of
96 phenotypic and transcriptomic stages [18]. This spectrum makes the assignment of identified
97 cell types to known cell types challenging. In our original MATLAB version, the cell-type
98 assignment was based on correlation analysis using a user-provided pure cell line GEP. While
99 this option remains available, the increasing number of publicly available single cell data
100 provide alternative GEPs for cell-type assignment. Thus, in CDSeqR, we designed and
101 implemented a new function to perform cell-type assignment for the downstream processing of
102 cell-type-specific GEPs. The cell-type assignment function takes advantage of publicly available
103 scRNA-seq data such as in [19,20].
104
105 Specifically, we generated the synthetic scRNA-seq data in two steps. First, under the
106 assumption that the total read count of each CDSeq-estimated cell type followed a negative
107 binomial distribution, we randomly generated the total synthetic read count for each cell type
108 using the R function *rnbinom (n, size, mu)* with *n* set equal to 1 and, by default, *mu* set equal to
109 the maximum of all single-cell read counts in the single-cell data set and *size* set equal to mu^2

110 (so that the negative binomial variance was $mu + mu^2/size = mu + 1$). These default parameters
111 in *rnbinom* can be replaced by user-provided parameters. Second, we generated the synthetic
112 single-cell read counts of every gene for each CDSeq-estimated cell type based its
113 corresponding CDSeq-estimated gene expression profile using the R function *rmultinom(n, size,*
114 *prob)*, with *n = 1*, *size* set equal to the total read count for the cell type as generated by the
115 preceding call to *rnbinom*, and *prob* provided by CDSeq-estimated cell-type-specific gene
116 expression profile expressed as a multinomial parameter vector (i.e., expression levels
117 expressed as proportions that sum to one). The resulting synthetic scRNA-seq data are
118 subsequently combined with publicly available scRNA-seq data.
119
120 We then performed clustering analysis on the combined dataset using the Seurat package
121 [21,22]. The clustering algorithm implemented in Seurat is a graph-based approach based on
122 the outcomes of principal components analysis. The number of clusters cannot be set explicitly
123 but is affected by a resolution parameter; the larger the resolution parameter is the more
124 clusters will be returned. The cell types identified by CDSeq from bulk RNA-seq data were
125 clustered with the individual cells from the scRNA-seq data. We then assigned CDSeq-identified
126 cell types to the cell type from the scRNA-seq data to which the majority of the cells in the
127 cluster belong (Fig 1).
128

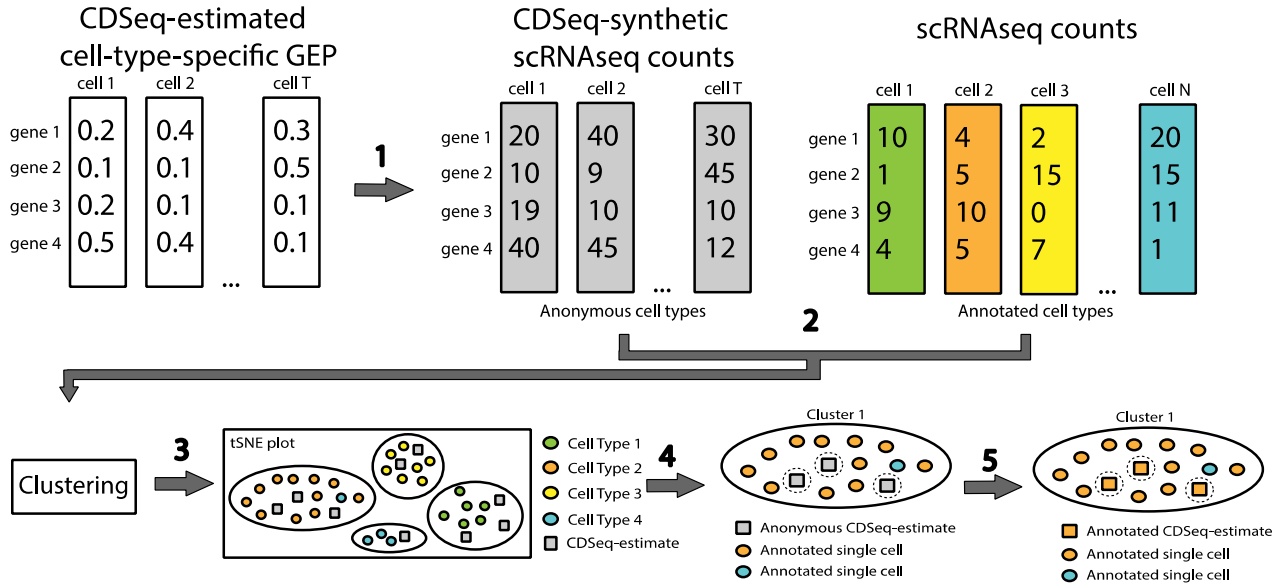
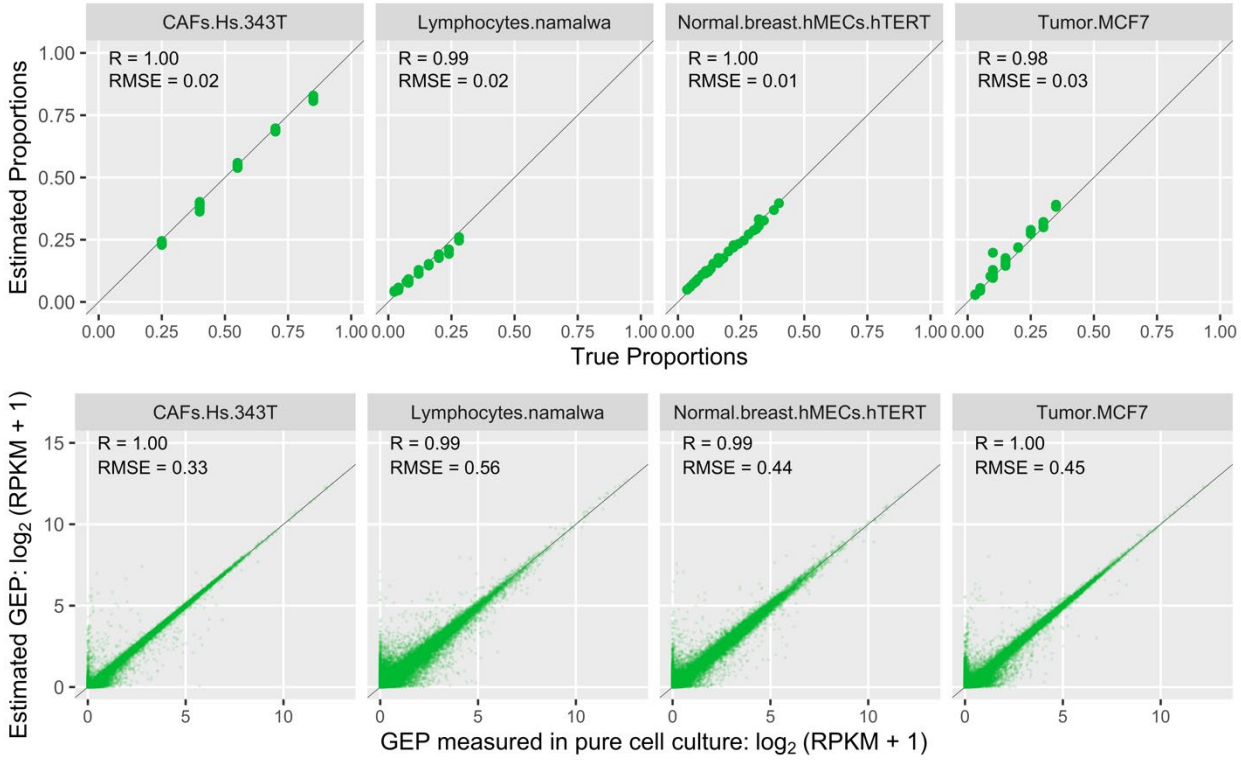


Figure 1. Schematic illustration of our cell-type assignment approach. It consists of five steps: 1) generating synthetic single-cell read counts (gray rectangles) for the CDSeq-estimated cell-type-specific GEPs (white rectangles); 2) combining the pseudo scRNA-seq data with experimental scRNA-seq data with cell annotation (colored rectangles), then performing clustering using Seurat; 3) visualizing the result using the tSNE plot to show the 2D embedding of the clustering outcome with differently colored circles denoting distinctive cell types in the single-cell data and gray squares denoting the CDSeq-identified cell types; 4) annotating each cluster to the cell type to which the majority of the cells in the cluster belong; 5) assigning the CDSeq-identified cell types accordingly as in step 4.

Results

We validated CDSeqR using both *in silico* and *in vitro* mixtures [9] where both cell-type composition and cell-type-specific GEPs are known. For *in vitro* mixtures, the CDSeqR estimates

144 matched the known cell-type-specific GEPs and sample-specific cell-type proportions (Fig. 2)
145 while reducing computational time by about 96% (from about 120 min to 5 min). In our
146 previous work, we compared CDSeq with seven other deconvolution methods and illustrated its
147 advantages [9]. Here, we further compared CDSeqR with a recently developed complete
148 deconvolution method, linseed [10], that was not part of our previous comparison. CDSeqR
149 software and linseed both provided accurate cell-type proportion using similar amounts of
150 computer time; however, linseed cannot estimate cell-type-specific GEPs (Supplementary Fig.
151 S1-S5). We further tested CDSeqR using TCGA data and showed that CDSeqR-identified cell
152 types are consistent with single-cell RNA-Seq results (Puram et al., 2017) (Supplementary Fig.
153 S5).



154

155

156 Figure 2. Results of CDSeqR analysis of the *in vitro* cell mixtures. We ran CDSeqR on 32 bulk
157 RNA-seq samples with known cell proportions in each sample [23] using 7 randomly sampled
158 gene subsets of size 500 (7 data blocks, each of 500 genes). CDSeqR finished within 5 minutes
159 on an iMAC with 3.5GHz Intel Core i7 processor whereas the original CDSeq (MATLAB version)
160 took about 2 hours. In the first row, each point represents a sample (n=32); in the second row,
161 each point represents a gene (n=~19,000). True and estimated GEPs are normalized as RPKM.
162 The line in each panel is the x=y line.
163
164 One input parameter that CDSeq uses is the number of cell types to be deconvolved. We
165 suggest setting a large number of cell types when deconvolving real bulk RNA-seq samples. We
166 found that, when the number of cell types is set too small, the algorithm may fail to capture the
167 major cell types due to heterogeneity within those cell types [18]. The input for the appropriate
168 number of cell types is data-dependent. More tissue samples may require a larger number of
169 cell types to allow CDSeq to detect all the major cell types. The implementation allows one to
170 set the number of cell types to be a vector. Then, CDSeqR will perform computation in parallel
171 for each value in the vector and see which value yields the cell types that are in a better
172 agreement with the single-cell data. As an example, we showcased the usage of deconvolution
173 using CDSeqR and our new cell-type annotation function by deconvolving the GTEx
174 (www.gtexportal.org) V8 frontal cortex data (209 samples). We also used publicly available
175 brain single cell RNA-seq data to assign the CDSeq-estimated cell types [20]. The result
176 demonstrates that setting the number of cell types to 200 enabled detection of the major cell
177 types in heterogenous brain samples (Fig. 3). Although 200 may seem large, it may be justifiable

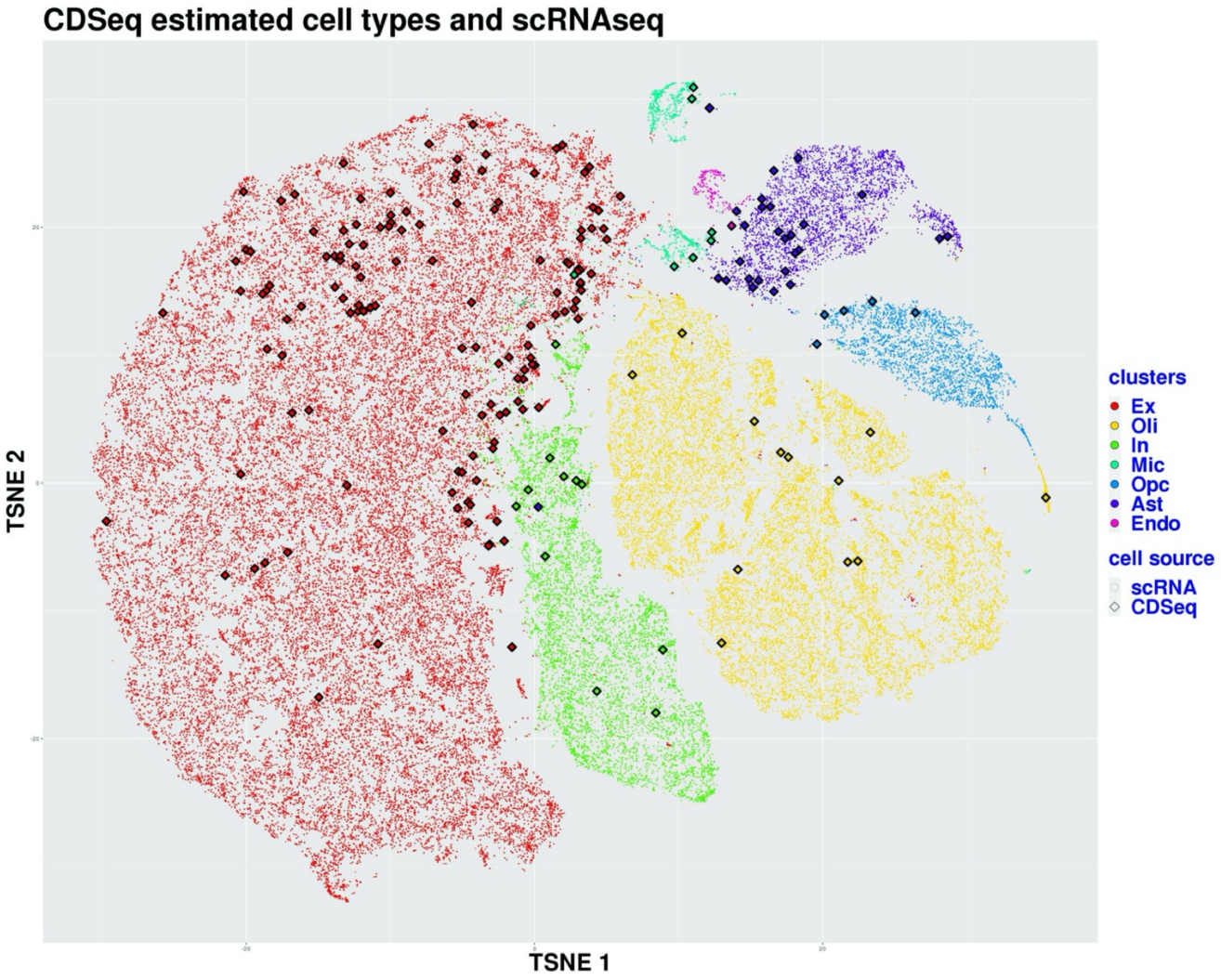
178 because the brain is highly heterogenous and cell states are known to represent a continuum.

179 This notion is consistent with the finding that, in brain prefrontal cortex [20,24], many of the

180 major cell types consist of multiple subtypes. For example, excitatory neurons have more than

181 ten subtypes [20,24].

182



183

184 Fig. 3. tSNE plot of the CDSeq-identified cell types in 209 GTEx frontal cortex RNA-seq bulk

185 tissue samples and annotated individual cells from scRNA-seq data [20]. In the tSNE plot, the

186 colored dots represent single cells of brain prefrontal cortex tissues from 48 individuals [20].

187 The colored squares denote the 200 CDSeq-identified cell types. As shown, CDSeq-estimated
188 cell types overlapped with all major clusters of single cells. CDSeqR identified the following cell
189 types: excitatory neurons (Ex), inhibitory neurons (In), Oligodendrocyte (Oli), microglia (mic),
190 oligodendrocyte progenitor cell (opc), astrocytes (ast), and endothelia (endo).
191
192 **Discussion and Conclusions**
193 CDSeq is a complete deconvolution method that uses gene expression data alone to estimate
194 both sample-specific cell-type proportions and cell-type-specific gene expression profiles [9].
195 CDSeq was originally developed in MATLAB. In this work, we developed an R package, CDSeqR,
196 for a broader group of users. Moreover, in the R implementation, we substantially improved
197 CDSeq's computational efficiency in both speed and memory usage. Specifically, we proposed a
198 novel strategy – Reduce and Recover – to allow CDSeq to run on multiple processors
199 simultaneously (parallelization). We further validated CDSeqR package using both *in silico* and
200 *in vitro* cell mixture data. We compared CDSeqR performance with another recently developed
201 complete deconvolution method and showed that, using similar amounts of computational
202 time, cell-type proportions estimated by CDSeqR are as accurate as those estimated by linseed;
203 but CDSeqR also provides cell-type-specific GEP estimates, whereas linseed does not.
204
205 In the R package, we also implemented a new cell-type assignment function. This new function
206 utilizes publicly available scRNA-seq datasets to help annotate and visualize CDSeq-estimated
207 cell types. In our R package, we provided publicly available scRNA-seq datasets for the major
208 human organs from the human cell landscape [19]; single cells in these data sets are already

209 annotated as to cell type. We transform the CDSeqR estimated cell-type-specific gene
210 expression profiles into “single-cell read counts” (synthetic scRNA-seq data). We then combine
211 our synthetic scRNA-seq data with experimentally obtained scRNA-seq as a single data matrix
212 for clustering. We identify the CDSeqR estimated cell types within a cluster by the cell type of
213 majority of the previously annotated single cells in that cluster.
214
215 As single-cell studies become popular, the complex nature of cell types is more clearly
216 recognized as new cell subtypes continue to be discovered [18,20,25]. Those findings suggest
217 that cell types may be better described as a continuum rather than as a sequence of discrete
218 states. This recognition, consequently, brings challenges to deconvolution algorithms, especially
219 for those that rely reference gene expression profiles [4,8,26,27]. The references will not fully
220 capture the heterogeneity and the continuum of states exhibited by cell types in bulk tissues.
221 Compared to existing deconvolution methods, CDSeq can mitigate such issues of heterogeneity
222 owing to its unsupervised framework (does not require a reference panel). We provide a guide
223 for users to assess the number of cell types input to CDSeq for identifying the major cell types
224 and their associated subtypes.
225
226 Since cell types represent a continuum, it is difficult to assign a value to the number of cell
227 types present. This issue is analogous to determining the number of clusters from a clustering
228 analysis. Nonetheless, we suggest that, in practice, one may set the number of cell types to be a
229 vector that covers a range of plausible values. For example, set parameter cell_type_number =
230 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.

Complete deconvolution presents a different challenge – the assignment of biological identities to the nominal cell types identified by the algorithm. CDSeqR employs a new approach to deal with this challenge. Our approach takes advantage of the large amount of existing scRNA-seq data. By converting CDSeqR-estimated gene expression profiles to “single cell read counts”, one can easily assign the identified cell types to existing known cell types based on publicly available single-cell data.

In summary, we believe that our CDSeqR package will be a valuable tool for computationally deciphering sample heterogeneity using bulk RNA-Seq data.

Availability and requirements

Project name: CDSeq R package

Project home page: https://github.com/kkang7/CDSeq_R_Package

Operating system: Platform independent

Programming language: R, C++, C.

Other requirements: None

License: GPL-3

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.

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).

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

275 authors contributed to the writing of the manuscript. LL guided the research. All authors read
276 and approved the final manuscript.
277
278 **Acknowledgements**
279 The authors thank the NIEHS Office of Scientific Computing for computing support and the
280 NIEHS Epigenomics and DNA Sequencing Core Facility for RNA-seq analysis of the cell mixture
281 samples.
282
283 **Reference**
284 1. Demaria O, Cornen S, Daëron M, Morel Y, Medzhitov R, Vivier E. Harnessing innate
285 immunity in cancer therapy. Nature. 2019.
286 2. Zheng C, Zheng L, Yoo JK, Guo H, Zhang Y, Guo X, et al. Landscape of Infiltrating T Cells in
287 Liver Cancer Revealed by Single-Cell Sequencing. Cell. 2017;
288 3. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, et al. Cell type--
289 specific gene expression differences in complex tissues. Nat Methods. 2010;7(4):287–9.
290 4. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining
291 cell type abundance and expression from bulk tissues with digital cytometry. Nat
292 Biotechnol. 2019;
293 5. Li Y, Xie X. A mixture model for expression deconvolution from RNA-seq in
294 heterogeneous tissues. BMC Bioinformatics. 2013;14(5):S11.
295 6. Ahn J, Yuan Y, Parmigiani G, Suraokar MB, Diao L, Wistuba II, et al. DeMix: Deconvolution
296 for mixed cancer transcriptomes using raw measured data. Bioinformatics. 2013;

297 7. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of
298 cell subsets from tissue expression profiles. Nat Methods. 2015;
299 8. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-
300 subject single-cell expression reference. Nat Commun. 2019;
301 9. Kang K, Przytycka TM, Meng Q, Shats I, Umbach DM, Li M, et al. CDSeq: A novel complete
302 deconvolution method for dissecting heterogeneous samples using gene expression
303 data. PLOS Comput Biol [Internet]. 2019 Dec 2; Available from:
304 https://doi.org/10.1371/journal.pcbi.1007510
305 10. Zaitsev K, Bambouskova M, Swain A, Artyomov MN. Complete deconvolution of cellular
306 mixtures based on linearity of transcriptional signatures. Nat Commun. 2019;
307 11. Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific
308 information from heterogeneous samples. Curr Opin Immunol. 2013;25(5):571–8.
309 12. Erkkilä T, Lehmusvaara S, Ruusuvuori P, Visakorpi T, Shmulevich I, Lähdesmäki H.
310 Probabilistic analysis of gene expression measurements from heterogeneous tissues.
311 Bioinformatics. 2010;
312 13. Qiao W, Quon G, Csaszar E, Yu M, Morris Q, Zandstra PW. PERT: a method for expression
313 deconvolution of human blood samples from varied microenvironmental and
314 developmental conditions. PLoS Comput Biol. 2012;8(12):e1002838.
315 14. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of
316 heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics.
317 2013;29(8):1083–5.
318 15. Zhong Y, Wan YW, Pang K, Chow LML, Liu Z. Digital sorting of complex tissues for cell

319 type-specific gene expression profiles. BMC Bioinformatics. 2013;
320 16. Gaujoux R, Seoighe C. Semi-supervised Nonnegative Matrix Factorization for gene
321 expression deconvolution: a case study. Infect Genet Evol. 2012;12(5):913–21.
322 17. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci. 2004;
323 18. Pastushenko I, Brisebarre A, Sifrim A, Fioramonti M, Revenco T, Boumahdi S, et al.
324 Identification of the tumour transition states occurring during EMT. Nature. 2018;
325 19. Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, et al. Construction of a human cell landscape
326 at single-cell level. Nature. 2020;
327 20. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell
328 transcriptomic analysis of Alzheimer's disease. Nature. 2019;
329 21. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic
330 data across different conditions, technologies, and species. Nat Biotechnol. 2018;
331 22. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al.
332 Comprehensive Integration of Single-Cell Data. Cell. 2019;
333 23. Kang K, Meng Q, Shats I, Umbach DM, Li M, Li Y, et al. CDSeq: A novel complete
334 deconvolution method for dissecting heterogeneous samples using gene expression
335 data. PLoS Comput Biol. 2019;
336 24. Gandal MJ, Zhang P, Hadjimichael E, Walker RL, Chen C, Liu S, et al. Transcriptome-wide
337 isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science (80-).
338 2018;
339 25. Trapnell C. Defining cell types and states with single-cell genomics. Genome Res.
340 2015;25(10):1491–8.

341 26. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of
342 cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453.
343 27. Hunt GJ, Freytag S, Bahlo M, Gagnon-Bartsch JA. Dtangle: Accurate and robust cell type
344 deconvolution. Bioinformatics. 2019;
345