scientific reports

OPEN

Characterizing the composition of iPSC derived cells from bulk transcriptomics data with CellMap

Zhengyu Ouyang[1], Nathanael Bourgeois-Tchir[2], Eugenia Lyashenko[2,3], Paige E. Cundiff[2], Patrick F. Cullen[2], Ravi Challa[2], Kejie Li[2], Xinmin Zhang[1], Fergal Casey[2], Sandra J. Engle[2], Baohong Zhang[2] & Maria I. Zavodszky[2]✉

Induced pluripotent stem cell (iPSC) derived cell types are increasingly employed as in vitro model systems for drug discovery. For these studies to be meaningful, it is important to understand the reproducibility of the iPSC-derived cultures and their similarity to equivalent endogenous cell types. Single-cell and single-nucleus RNA sequencing (RNA-seq) are useful to gain such understanding, but they are expensive and time consuming, while bulk RNA-seq data can be generated quicker and at lower cost. In silico cell type decomposition is an efficient, inexpensive, and convenient alternative that can leverage bulk RNA-seq to derive more fine-grained information about these cultures. We developed CellMap, a computational tool that derives cell type profiles from publicly available single-cell and single-nucleus datasets to infer cell types in bulk RNA-seq data from iPSC-derived cell lines.

Human induced pluripotent stem cells (iPSC) and cells developed from them are gaining widespread acceptance for their use in understanding normal and disease processes in vitro[1,2]. Derived directly from patient material with an unlimited capacity to proliferate in the undifferentiated state and the theoretical ability to differentiate into any cell type under appropriate experimental conditions, iPSCs and their differentiated derivatives represent an attractive alternative to traditional in vitro models that rely on cancer cell lines or rodent material. Barriers that once limited the use of iPSCs, such as costly reagents, complicated culture protocols, or restricted access to high quality, well-characterized iPSC lines, have diminished over the last decade. It has also become increasingly apparent that human biology can diverge significantly from rodent and even non-human primate biology, thus necessitating the use of human cells[3,4].

For in vitro studies using iPSC-derived cells to be meaningful, it is important to understand the relationship of the in vitro generated cells to endogenous cell types. This is particularly true during the development of novel iPSC differentiation protocols when seemingly small changes in culture conditions can lead to divergent cell fates. Evaluating expression of only a few canonical proteins by immunocytochemistry or genes by qPCR may not provide an adequate representation of what is in the cell culture dish, and it can be technically challenging by these methods to test for the presence of off-target cell fates. A similar issue arises once a differentiation protocol has been established and one would like to ensure the reproducibility of specific cell type production within the same iPSC line and across different iPSC lines differentiated with the same protocol. While iPSC-derived cell types need not be perfect replicas of the endogenous cell type to be useful for disease modeling, they must reproduce the mechanism of the endogenous function to be evaluated in an appropriate cellular context to be physiologically relevant.

A variety of cell types exist in biological tissues performing different functions. When a biological function is altered or deficient, we need to understand the origin and mechanism of this aberration to devise ways of correcting it. However, most of the readily available biological samples are composed of a mixture of cell types. Experimentally separating these cell types and performing single-cell sequencing on them is cumbersome and costly. Therefore, developing computational approaches for cell type deconvolution from bulk RNA-seq data has been a popular and fruitful endeavor during the past decade[5,6]. The earlier methods relied on cell type marker genes or cell-specific signatures obtained from prior publications or derived from low-throughput experiments, such as gene expression profiling on FACS sorted cells. CellMix was one such useful toolset that provided access to multiple deconvolution methods allowing the user to select the best approach based on the available data[7]. Unfortunately, the support of this public tool has been discontinued and it is incompatible with the latest R

[1]BioInfoRx, Inc., 510 Charmany Dr, Suite 275A, Madison, WI 53719, USA. [2]Translational Biology, Research and Development, Biogen, Inc., 225 Binney St, Cambridge, MA 02142, USA. [3]Present address: Genomic Medicine Unit, Sanofi, 225 2nd Ave, Waltham, MA 02451, USA. ✉email: maria.zavodszky@biogen.com

libraries. More complex and computationally demanding approaches have also been designed to characterize engineered cells based on inference of tissue-specific gene regulatory networks first from microarray data and then from bulk RNA sequencing data[8,9].

The aspiration to eliminate the dependence on prior knowledge of cell type markers or expression signatures led to efforts to develop de novo deconvolution algorithms. One such example is CellDistinguisher, which mathematically identifies cell type specific patterns in bulk expression data obtained from multiple heterogeneous samples[10]. CellDistinguisher can identify the genes that best distinguish a defined number of cell types or biological processes in the input data. This type of unsupervised deconvolution, however, cannot tell whether the patterns detected are a result of the presence of multiple cell types or subpopulations undergoing different biological processes, e.g., apoptosis, different phases of the cell cycle, etc. Adding even limited amount of prior knowledge in the form of marker gene sets can guide such methods into the desired direction.

With the advances of single-cell and single-nucleus (sc/sn) RNA sequencing, generating the much-needed prior knowledge to characterize cell types is gradually becoming a reality[11,12]. These sequencing approaches are still quite expensive and arduous. Consequently, most laboratories cannot afford to apply them regularly to characterize iPSCs. But, given even a limited number of such datasets for the cell types of interest, they can be leveraged to mathematically characterize many samples for which the much easier and more cost-effective bulk RNA-seq data can be generated.

There are multiple deconvolution methods that rely on single-cell information to generate cell type expression profiles. Providing a thorough review of the field is beyond the scope of this work. Here we focus on a few representative examples that we used to benchmark our method. MuSiC uses scRNA-seq data to estimate cell type composition in samples with bulk RNA-seq data[13]. One limitation of this approach is that the query bulk needs to have the same cell types as the samples from which scRNA-seq data is derived. A more practical method would allow cell type profiles to be derived from sc/sn RNA-seq datasets of samples with disparate compositions of cell types. Bisque is designed to perform a very fast decomposition using non-negative least squares (NNLS) regression based on one single-cell dataset serving as a reference[14]. The constraint with this method, like with MuSiC, is that the user should have a single-cell dataset with cell types matching the bulk samples available before applying it. The most recent deconvolution method, SCDC derives expression profiles from multiple scRNA-seq datasets adopting an ensemble framework to implicitly address the batch effects inherent in datasets coming from different sources[15]. It achieves this by applying different weights to different datasets. The reference data most similar to the bulk data overall will have a higher weight. A drawback of this method is that all reference datasets need to have the same cell types. Meeting this condition is unlikely when using publicly available datasets from multiple sources.

For accurate disease modeling using human iPSCs, we needed a deconvolution tool designed to overcome the limitations encountered in existing tools. The new tool should be able to characterize a variety of cell types while making it possible to expand the list as needed. It should easily incorporate new reference datasets as they become available, allowing the user to retrain and retest the tool without modifying the code. To mitigate the batch effects or other biases inherent in individual datasets, we imposed a requirement to have each cell type represented in at least three datasets, but we do not impose a constraint that each reference dataset need contain all cell types of interest.

CellMap was developed to meet these requirements. It was aimed mainly at characterizing iPSC-derived cells in terms of their cell type composition, their similarity to previously characterized primary cells or other iPSC-derived datasets, as well as assessing batch-to-batch variability. One important feature of this tool is the ability to regenerate the cell type profiles easily as newer or better-quality single-cell and single-nucleus datasets are produced and published. Here we demonstrate that besides iPSC-derived cells, CellMap can naturally be applied to the deconvolution of any complex samples whose constituent cell types are represented in the provided reference datasets.

Methods

Workflow.

CellMap employs NNLS regression to decompose a bulk sample into cell type proportions based on the gene expression values (G) of a query bulk sample and the cell type specific expression profiles as shown in Eq. (1):

$$G = \beta_1 \cdot P_1 + \beta_2 \cdot P_2 + \cdots + \beta_c \cdot P_c \tag{1}$$

where G is a $g \times 1$ query expression vector with g denoting the number of genes. $P_1, P_2, \ldots, P_c$ are expression profiles (also $g \times 1$ vectors) of c pure cell types derived from single cell or single-nucleus RNA-seq datasets, while $\beta_1, \beta_2, \ldots, \beta_c$ are the derived non-negative proportions of these cell types in the bulk sample. The overall workflow consists of two major steps: generating the cell type profiles and deconvolution of the query bulk sample (Fig. 1). While deriving the cell type profiles is the more computationally expensive part, it only needs to be carried out once for a given set of input sc/sn datasets.

Generating pseudo-bulk samples.

Both pure cell-type specific and mixed pseudo-bulk samples were generated from sn/sc RNA-seq datasets. The pure pseudo-bulks were then subject to normalization to alleviate batch effects, followed by differential expression analysis to identify cell type specific genes that became part of the cell type profiles. We created three categories of profiles listed in order of decreasing numbers of cell types and increasing similarity between them: Major9, CNS6 and Neuron3. This was done to enable a hierarchical deconvolution approach, going from coarse-grained to more fine-grained deconvolution (more details under Stepwise deconvolution). The mixed pseudo-bulk samples were used as part of training process to select the top

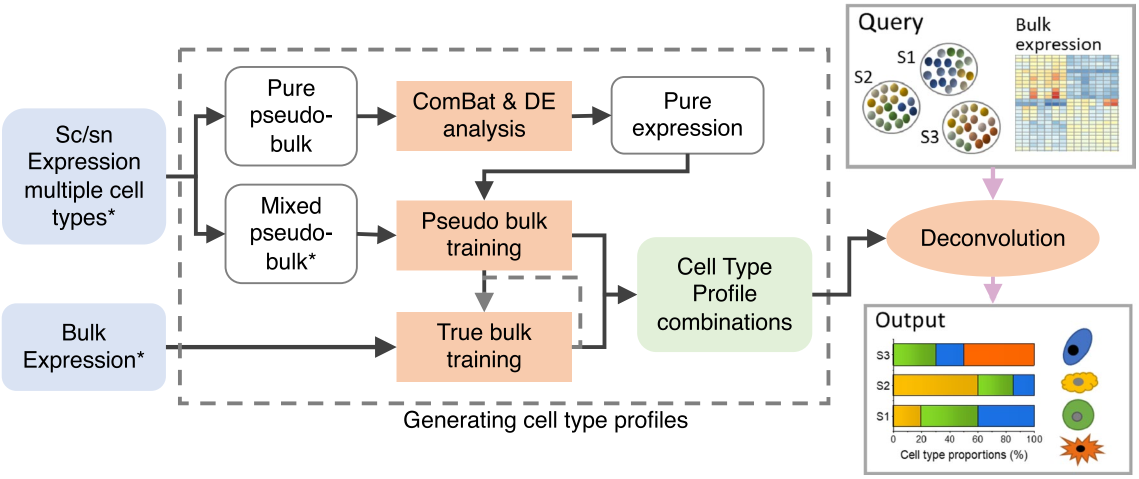
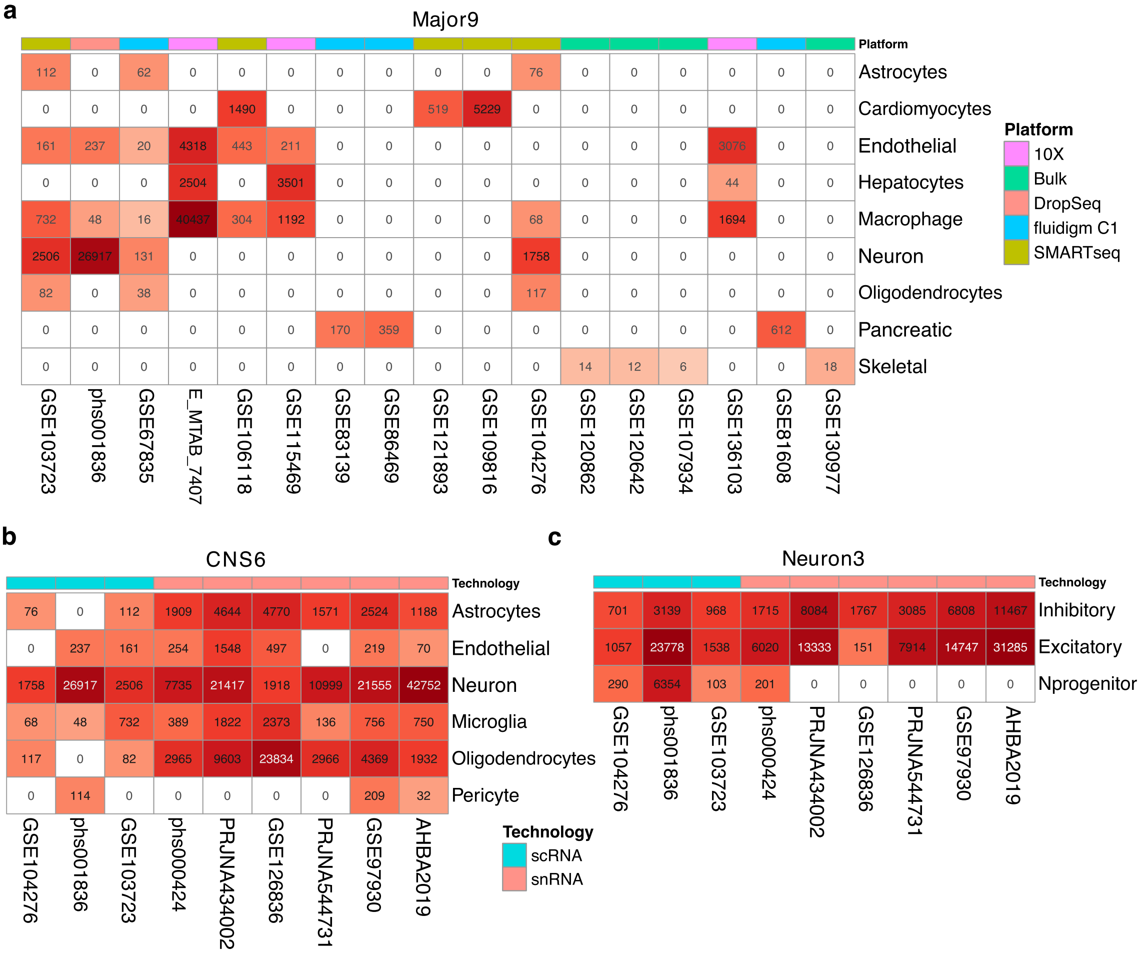


Figure 1. Overview of the CellMap workflow. CellMap includes the steps to generate the cell type specific profiles and the deconvolution which uses these profiles to predict the composition of a bulk sample. Datasets with samples of known cell-type compositions are denoted by *.



	GSE103723	phs001836	GSE67835	E_MTAB_7407	GSE106118	GSE115469	GSE83139	GSE86469	GSE121893	GSE109816	GSE104276	GSE120862	GSE120642	GSE107934	GSE136103	GSE81608	GSE130977
Astrocytes	112	0	62	0	0	0	0	0	0	0	76	0	0	0	0	0	0
Cardiomyocytes	0	0	0	0	1490	0	0	0	519	5229	0	0	0	0	0	0	0
Endothelial	161	237	20	4318	443	211	0	0	0	0	0	0	0	0	3076	0	0
Hepatocytes	0	0	0	2504	0	3501	0	0	0	0	0	0	0	0	44	0	0
Macrophage	732	48	16	40437	304	1192	0	0	0	0	68	0	0	0	1694	0	0
Neuron	2506	26917	131	0	0	0	0	0	0	0	1758	0	0	0	0	0	0
Oligodendrocytes	82	0	38	0	0	0	0	0	0	0	117	0	0	0	0	0	0
Pancreatic	0	0	0	0	0	0	170	359	0	0	0	0	0	0	0	612	0
Skeletal	0	0	0	0	0	0	0	0	0	0	0	14	12	6	0	0	18

	GSE104276	phs001836	GSE103723	phs000424	PRJNA434002	GSE126836	PRJNA544731	GSE97930	AHBA2019
Astrocytes	76	0	112	1909	4644	4770	1571	2524	1188
Endothelial	0	237	161	254	1548	497	0	219	70
Neuron	1758	26917	2506	7735	21417	1918	10999	21555	42752
Microglia	68	48	732	389	1822	2373	136	756	750
Oligodendrocytes	117	0	82	2965	9603	23834	2966	4369	1932
Pericyte	0	114	0	0	0	0	0	209	32

	GSE104276	phs001836	GSE103723	phs000424	PRJNA434002	GSE126836	PRJNA544731	GSE97930	AHBA2019
Inhibitory	701	3139	968	1715	8084	1767	3085	6808	11467
Excitatory	1057	23778	1538	6020	13333	151	7914	14747	31285
Nprogenitor	290	6354	103	201	0	0	0	0	0

Figure 2. Single-cell and single-nuclei datasets used for cell type profile generation. (**a**) Datasets used for the 9 major cell types. (**b**) Datasets used for the 6 CNS cell types. (**c**) Datasets used for the neuronal subtypes of interest. The numbers in the colored boxes indicate the number of cells contained in each category after curation. Inkscape v1.0.2 (https://inkscape.org/) was used to assemble the heatmaps.

performing cell type profile sets of 9, 6 and 3 cell types for the Major9, CNS6 and Neuon3 profiles, respectively (Fig. 2).

The cell type expression profiles were derived from publicly available sc/sn RNA-seq datasets with cell type annotations provided by the authors (Supplementary Table S1). For skeletal muscle profiles, we could not find human single cell RNA-seq datasets with adequate representation of the mature multinucleated myofibers. The available single cell datasets tend to be dominated by the mononucleated fraction of skeletal muscle (immune, endothelial, satellite and other mononuclear cells) with only a minor fraction of myofibers captured. Because of this, we decided to use bulk RNA-seq data obtained from isolated muscle fibers or muscle biopsy samples as if they were pseudo-bulk samples generated from single cells to derive the skeletal muscle profiles.

Since no two cells are identical, even if they are labeled as the same type in a single cell dataset, we attempted to capture this inherent biological variability by generating multiple pure cell type pseudo-bulk samples from each dataset. Cells were randomly selected from the existing pool of a given type within each dataset (not mixing input datasets) and their expression values were summed up across the cells for each expressed gene. Each of these random subsets contained at least 50% of the cells from the original dataset and at least 2 million total counts. Similarly, mixed cell type pseudo-bulk samples with known cell type compositions were also generated from each dataset for the profile selection training. After the pseudo-bulk samples were normalized to 1 M reads, genes with low expression were filtered out. Low expression genes were defined as having less than 4 counts in more than 20% of pseudo-bulk samples of any cell type. The exclusion of the low expression genes resulted in losing 10% or less of the total read counts in all but one of the datasets used for the Major9 profiles. In the case of the datasets used for the profiles of CNS6 and Neuron3, this filtering removed less than 20% of the total read count in all but one dataset. The number of genes remaining after filtering was 18,715 for Major9 and 18,046 for CNS6 and Neuron3 (Supplementary Fig. S1). The expression metric used for full length transcript datasets was TPM (Transcripts Per Million), while UMI (Unique Molecular Identifier) was used in the case of datasets obtained with 3′ RNA sequencing.

Normalization across datasets. ComBat normalization was used to eliminate batch effects among pure pseudo-bulk samples across reference datasets that can occur due to the differences in sequencing library preparation and sc/sn sequencing platforms[16]. Depending on the extent of overlapping cell types among datasets, two different strategies were applied. If shared cell types among datasets were sparse (Fig. 2a), ComBat was applied to the pseudo-bulk pure samples independently for each cell type across their source datasets. However, in the preferred scenario, when the overlap of cell types among the datasets was high (Fig. 2b,c), the normalization was performed across all pseudo-bulk samples generated for the pure cell types in one step. This approach was taken to avoid removing true cell type differences while performing batch correction for unbalanced datasets.

Cell type profile genes. To increase the sensitivity of cell type detection, only cell type specific genes were included in the cell type profiles. Such genes were more highly expressed in one cell type relative to all others and they were identified by performing pairwise differential expression analysis with edgeR on the batch-corrected pseudo-bulk data[17]. The specific cutoff values were log2FoldChange > 1 in one cell type versus the average expression in the others, and adjusted p-value < 0.05. For a balanced approach, an attempt was made to maintain a similar number of profile genes for all cell types within a group (see description of cell type groups under Stepwise Deconvolution below). Thus, fewer genes were kept for the CNS6 profiles than those of Major9, because the more similar cell types in CNS6 resulted in fewer differentially expressed genes. The number of genes in the Major9 profiles ranged from 205 to 351, while the CNS6 and Neuron3 profiles had 119 to169 genes (Supplementary Fig. S1).

Training. The goal of the training was to generate a collection of cell-type profiles used in the composition estimation procedure (Eq. 1). The training was performed in iterations until either the desired performance or the maximum number of iterations was reached. Two sequential training steps were taken in each iteration: (1) on mixed pseudo-bulk samples and (2) on real bulk samples, both with known compositions. A detailed description of the training protocol is provided in the Methods section of the Supplementary Material. Briefly, top performing sets of profiles in the mixed pseudo-bulk training were further ranked and filtered in the real bulk training. The performance was evaluated by the root mean square error (RMSE) between the expected and predicted proportions:

$$RMSE = \sqrt[2]{\frac{\sum_i^c (Exp_i - Est_i)^2}{c}} \tag{2}$$

where Exp_i and Est_i are the expected and estimated proportions of cell type i when the evaluation is done across a total number of c different cell types of a sample. At the end of the training, multiple sets of profiles were kept for deconvolution.

Deconvolution. Hundreds of profile sets were generated by combining the individual profiles (pure pseudo-bulks). These were ranked in the first training step on mixed pseudo-bulk samples, followed by further ranking and filtering on the real bulk training samples. During deconvolution, an ensemble method was adopted to integrate the deconvolution results from different profile sets for optimizing the match between the profiles and the query data. For each query bulk sample, the composition predictions of the top performing N profile sets were kept based on the goodness-of-fit RMSEs. The composition predictions of these top profile sets were then combined into weighted coefficients for each cell type:

$$\beta_i = \sum_{j}^{N} \beta_i^j \cdot w^j, \forall i = 1, 2, \ldots, c \tag{3}$$

where β_i^j is the proportion of i-th cell type in j-th estimated composition; the weight w^j was calculated based on the goodness-of-fit $RMSE_f$ of each set of profiles [i.e., $RMSE$ from observed and fitted values of gene expression in Eq. (1)]:

$$w^j = \frac{\left| RMSE_f^j - \overline{RMSE_f} \right|}{\sum_{k}^{N} \left| RMSE_f^k - \overline{RMSE_f} \right|}, \forall j = 1, 2, \ldots, N \tag{4}$$

Consequently, the profile set with the lowest RMSE had the largest contribution to the final composition. The results presented here were generated with N = 5, that is the cell type profiles with the lowest 5 RMSE values from all profile sets.

Stepwise deconvolution. Attempting to decompose samples into a very large number of cell types can be difficult as multiple small fractions are hard to predict accurately. To avoid this pitfall, we envisioned a multi-step process: in the first step, the major cell types of interest would be queried, followed by more refined deconvolution steps querying a narrower set of cell types or subtypes based on the outcome of the first step. Following this design, we curated two groups of datasets to include the cell types of interest. The first group consisted of 17 datasets to account for 9 major cell types (Major9), including astrocytes, cardiomyocytes, endothelial cells, hepatocytes, macrophages, neurons, oligodendrocytes, pancreatic and skeletal muscle cells. These are the main cell types of interest for neurodegenerative disease research as well as unintended types that might arise as off-target cell fates. The second group of 9 datasets focused on cell types specific for the central nervous system (CNS). From this dataset, we derived two sets of profiles: the CNS6 that included neurons, astrocytes, endothelial cells, microglia, oligodendrocytes and pericytes, as well as Neuron3 representing neuronal progenitors, inhibitory and excitatory neurons (Fig. 2).

Comparison to other deconvolution approaches. The mixed pseudo-bulk samples with known cell type proportions that were generated as part of the CellMap pipeline were also used to compare the performance of CellMap to three publicly available methods: MuSiC[13], SCDC[15] and Bisque[14].

Even though multiple sc/sn datasets were used for the profile generation, the input to the CellMap deconvolution step was a reduced data bundle comprised of the cell type profiles. The other three methods (MuSiC, SCDC and Bisque) work directly with sc/sn datasets instead of the pre-generated profiles. To avoid loading all input sc/sn datasets into the memory for the comparisons, an expression matrix was created by merging subsets of cells from them. At least half of the cells were randomly selected for each cell type from each dataset, not to exceed 20 M cells of a certain type. Three such input sc/sn expression matrices were created corresponding to the three CellMap profile sets.

The implementations of MuSiC, SCDC and BiSque were slightly modified from their published versions deposited in GitHub. The changes included adjustments to the output format and enabling parallel computing. The modified applications are also available from GitHub together with the expression matrices of the pseudo-bulk samples used for the benchmarking.

The comparison was performed for each cell type separately. The RMSE values were calculated across each cell type by comparing the expected and predicted compositions of the pure and mixed pseudo-bulk training samples. The RMSE was also computed for each pseudo-bulk sample across all cell types of a cell type group (Major9, CNS6, Neuron3).

Training datasets. The sc/sn RNA-seq datasets (bulk RNA-seq for muscle cells) used to generate the cell type profiles were also used for the pseudo-bulk training (Supplementary Table S1). Multiple rounds of random selections of cells minimized the overlap between the cells used for profile generation (from pseudo-bulk pure samples) and training (on pseudo-bulk mixed samples). Diseased samples were excluded when appropriate annotations were provided. For consistency, we reran the RNA-seq pipeline on these datasets when the raw data was available, as noted in Supplementary Table S1. In addition to these pseudo-bulk samples, true bulk samples from public repositories were used for the true bulk training.

Testing datasets. To evaluate the performance of CellMap, we assembled a collection of true bulk RNA-seq datasets of purified primary cells and iPSC-derived cell lines, independent from the training sets used in CellMap. For the datasets with primary cells, the information provided by the authors about the cell type composition of the samples was accepted to be the ground truth. In the case of the iPSC-derived cells, the entire target cell type was used as expected composition for the purpose of computing an RMSE. Even though these RMSE values reflected more on the deviation of cell line from the target cell type than on the performance of CellMap, they were deemed to be useful for detecting the changes in composition and similarity of iPSC derived cells relative to the primary cell types. In addition, we used bulk RNA-seq data from brain tissues of the ROSMAP dataset with matching immunohistochemistry (IHC) and snRNA-seq data as ground truth[18,19] (Supplementary Table S2).

Results

Normalization across datasets.

As revealed by principal component analysis (PCA), the pure pseudo bulk samples without any prior normalization tended to cluster by data sources rather than by cell types (Fig. 3a,b, Supplementary Figs. S2 and S3). This was not surprising, given the differences in sequencing library preparation and various sc/sn sequencing platforms employed to generate the datasets. This strong batch effect would have adversely affected the performance of the decomposition algorithm. Depending on the presence of overlapping cell types among datasets, two different ComBat normalization strategies were deployed to eliminate batch effects. The overlap of cell types among the reference datasets used to generate the Major9 cell type profiles was sparse (Fig. 2a). In this scenario, ComBat was applied to the pseudo-bulk pure samples for each cell type separately across their source datasets. The pseudo-bulk samples, that clustered by sequencing platforms originally, became clearly grouped by cell types after this batch correction (Fig. 3d). This strategy proved to be necessary because a normalization across all datasets and all cell types tended to remove cell type specific expression patterns (Fig. 3c). On the contrary, the datasets used for the CN6 and Neuron3 profiles had highly overlapping cell types (Fig. 2b,c). In this case, ComBat normalization across all pseudo-bulk samples in one step was sufficient to remove the batch effects without reducing signals due to real cell type differences. As shown on the corresponding PCA plots, the samples that originally grouped by datasets and sequencing technologies were properly clustered by cell types after normalization (Supplementary Figs. S2 and S3).

Following normalization, a subset of profile genes was selected for each cell type. Care was taken to maintain comparable overall expression profiles of these subsets across cell types, while assuring that the profile genes were expressed at higher levels in their respective cell type compared to all other types (Fig. 3e–g, Supplementary Figs. S2 and S3).

Performance on the training data.

Using the Major9 profiles, CellMap predicted the composition of the pseudo-bulk mixed samples with median RMSE values below 0.1 for all the 9 cell types (Fig. 4a). The lowest performance was observed on endothelial cells and macrophages. When computed across cell types in each dataset, the median RMSE was below 0.1 for the pseudo bulk samples generated from all but one dataset (Fig. 4b). The median RMSE was the highest for the true bulk samples (Fig. 4b last column). This latter is not surprising because, even though these true bulk samples contain purified cell types, they are rarely 100% pure and homogeneous.

The predictions using the CNS6 profiles were slightly less accurate on average. Having to combine single-cell and single-nuclei datasets for the CNS6 profile set increased the variability of the expression profiles. As a result, predicting the neuronal fractions accurately proved to be more challenging, with RMSE values reaching 0.2 (Fig. 4c). The prediction accuracy across datasets was relatively similar since they tended to have most of the CNS6 cell types: four datasets had median RMSE below 0.1 and five datasets had median RMSE between 0.1 and 0.2. The median RMSE of the predictions of the true bulk samples was well below 0.1, but the individual RMSE values fell into a wide range from 0.01 to slightly over 0.3 (Fig. 4d).

Predicting the neuronal subtypes in Neuron3 proved to be the most difficult task. The median RMSE values ranged from 0.02 to 0.15 for the three subtypes: inhibitory, excitatory and progenitors (Fig. 4e). On a dataset-by-dataset level, 4 datasets yielded accurate predictions with median RMSEs at or below 0.1, while the other 5 had higher RMSE values with larger spread (Fig. 4f). We attribute this increased difficulty to the often-subtle differences between the expression profiles of inhibitory and excitatory neurons that sometimes result in inconsistent labeling of these neuronal subtypes in different datasets. Cell type identification in scRNA-seq dataset clusters, whether it is done manually or by automated annotation tools, is a challenge because the clusters are not completely homogeneous. Slight differences in the selection of cluster marker genes can lead to discrepancies in cell type annotations, especially when these cell types are very similar to each-other, as pointed out by authors working on methods for automated annotation of cell types in sc/sn RNA-seq data[20,21]. Furthermore, the neuronal progenitors correspond to a continuum of cells at different stages of maturity which makes their characterization difficult. This is reflected in the variability of the prediction accuracy of these cell types across datasets (Fig. 4e).

Testing results.

The performance of CellMap was evaluated on independent bulk RNA-seq datasets by computing the RMSE between the predicted and expected compositions. We performed the tests with both sets of profiles that contained the cell types matching those expected to be found in the test datasets, the Major9 and CNS6. The calculated RMSE values were grouped by purified primary cell types, bulk samples from the ROSMAP dataset, and iPSC-derived cell types. Given a lack of quantitative information about the true cell type content, the expected composition of the iPSC-derived cells was set to 100% of the target cell type in order to allow the calculation of the RMSE.

On the primary cells, the performance of CellMap, as measured by RMSE, was comparable to the performance on the training datasets. Generally, higher RMSE values were obtained on the iPSC-derived cell lines, reflecting their imperfect resemblance to the target primary cell types. It might also be the case that such cell cultures include cells that are not fully differentiated or have stray fates. Median RMSE values were below 0.2 for the purified primary cell type samples using either set of profiles, except for pericytes, while the median RMSE of iPSC-derived samples were below 0.3, except for a subset of iPSC-derived astrocytes (Fig. 5a,b). The neurons and microglia (or macrophages in the case of Major9) were predicted close to the expected 100% purity. The astrocytes and pericytes proved to be the most challenging. The availability of reference purified pericyte datasets was limited. The ones we identified contained less than 100% pericyte-like cells and expressed genes also considered to be fibroblast and oligodendrocyte markers[22].

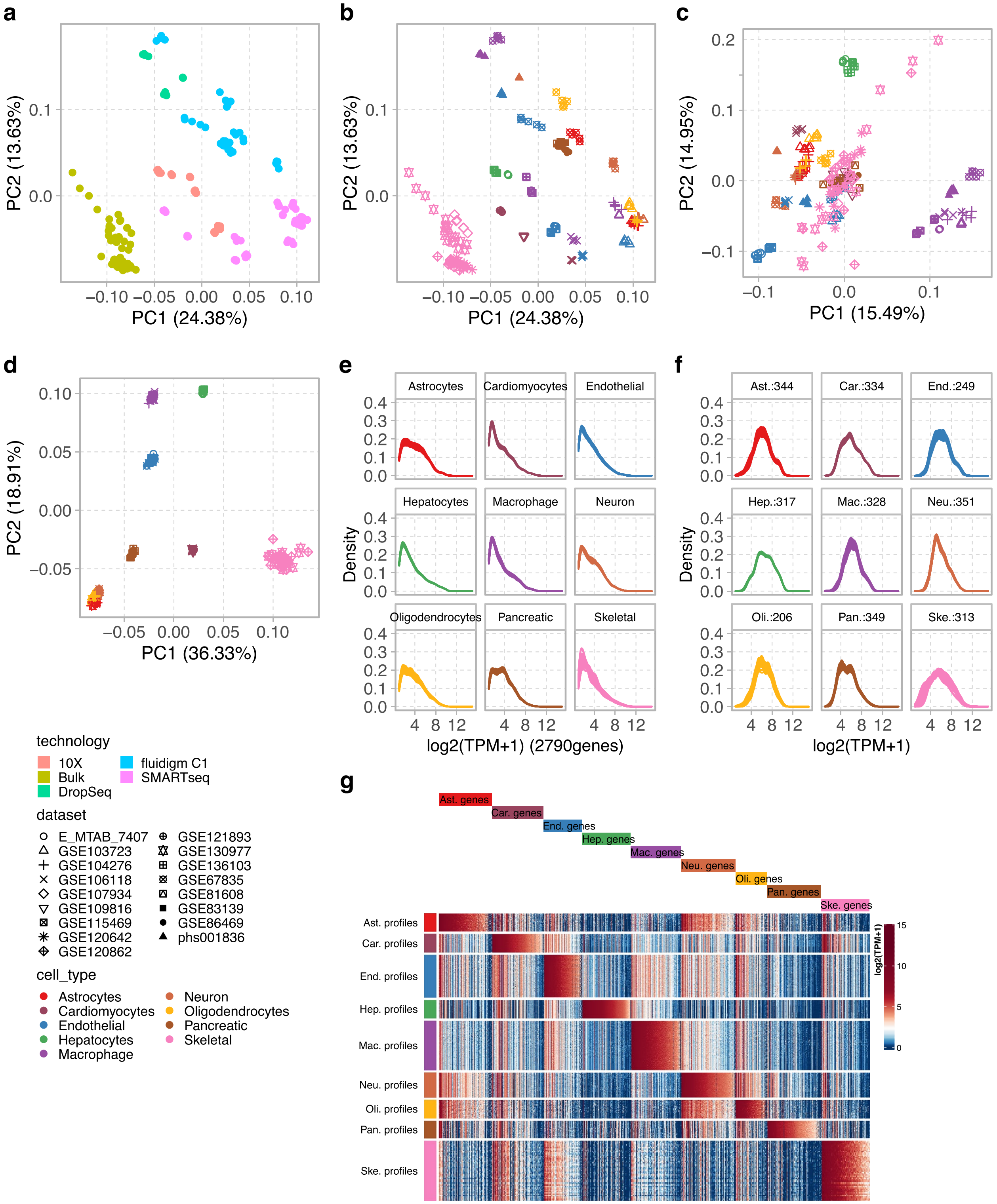


Figure 3. Major9 cell type profiles. (**a**,**b**) The first two principal components of the pseudo-bulk expression profiles of the Major9 cell types before normalization colored by sequencing platforms of the source datasets (**a**) and by cell types (**b**) showed the samples grouping by sequencing platforms. (**c**) After ComBat normalization across all datasets, cell type information was lost. (**d**) After ComBat normalization separately on each cell type, the pseudo-bulk samples grouped by cell types, as expected. (**e**) Expression profiles of all genes in the pure pseudo-bulk samples. (**f**) A subset of profile genes was selected such that their expression levels were comparable across cell types. The numbers next to the cell type labels indicate the number of genes in each of the cell type profiles. (**g**) The expression of the profile genes was higher in their respective cell type compared to all other types. Inkscape v1.0.2 (https://inkscape.org/) was used to assemble the figure.

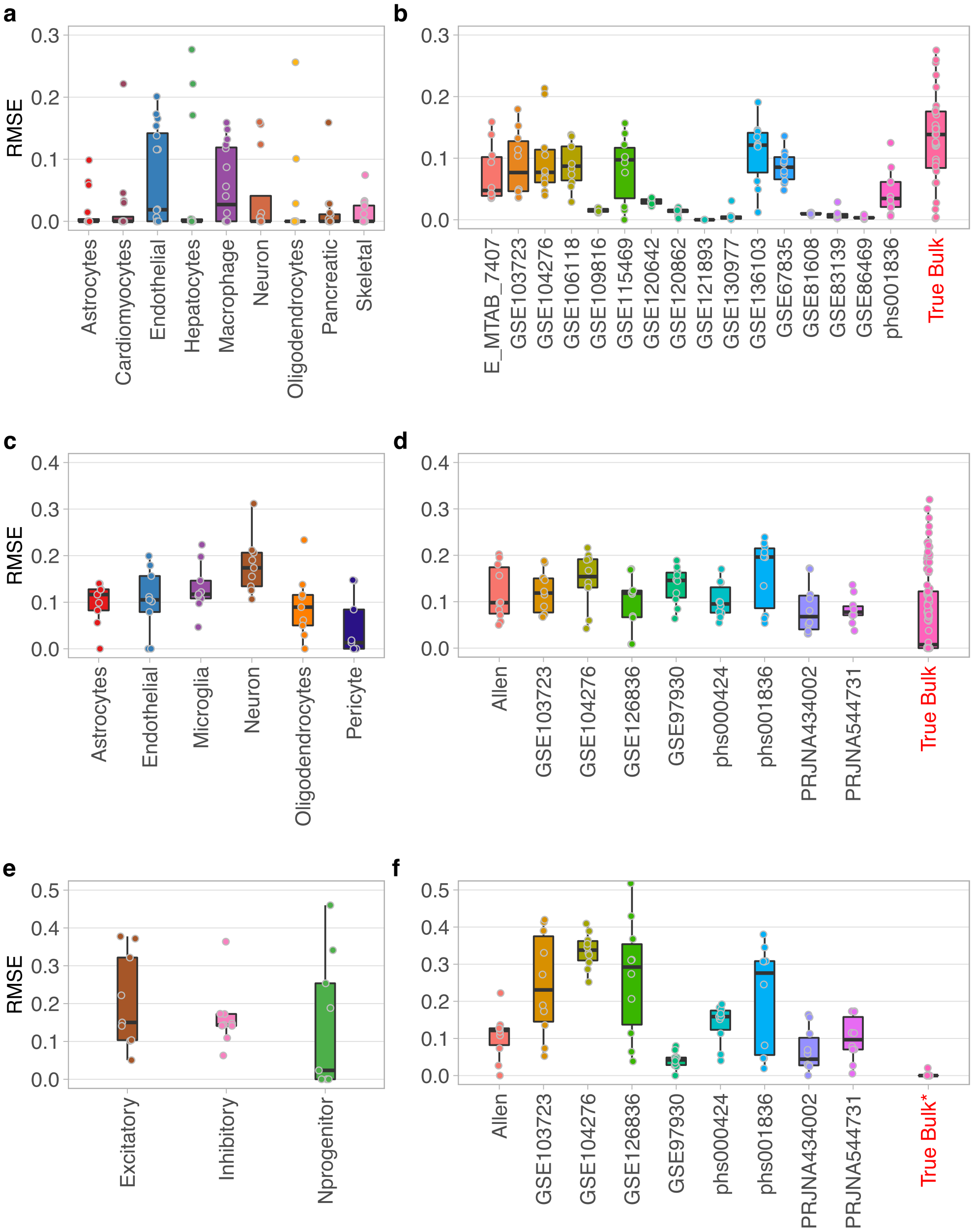


Figure 4. The performance of CellMap on pseudo-bulk mixed samples and true bulk using the three sets of profiles. (**a**,**b**) Datasets used for Major9 profiles; (**c**,**d**) for CNS6 profiles; (**e**,**f**) for Neuron3 profiles. (**a**,**c**,**e**) RMSE values by cell types with each data point representing one dataset; (**b**,**d**,**f**) RMSE values by test datasets with each data point representing one pseudo-bulk mixed or one true bulk sample.

The differences in performance with the Major9 and CNS6 profile sets confirmed that the input datasets used for generating the cell type expression profiles had a substantial influence on the outcome. This was also confirmed by the varying RMSE values grouped by datasets (Fig. 5c,d).

Examples of CellMap applications. The iPSC-derived cardiomyocyte dataset (GSE122380) was generated from a time course with 16 time points and 19 human cell lines capturing differentiation from iPSCs to mature cardiomyocytes[23]. Despite variable cardiomyocyte purity and marker gene expression levels across samples, CellMap clearly revealed the increasing cardiomyocyte fraction with time (Fig. 6a).

Microglia were derived from iPSCs in-house and the effects of different conditions and treatments during differentiation were tested. CellMap was applied to this dataset to determine the optimal time frame for differentiation and to assess how the treatments that such cell lines might have to be subjected to would alter the outcome. CellMap correctly indicated that the precursors were already microglia-like, and the microglia content

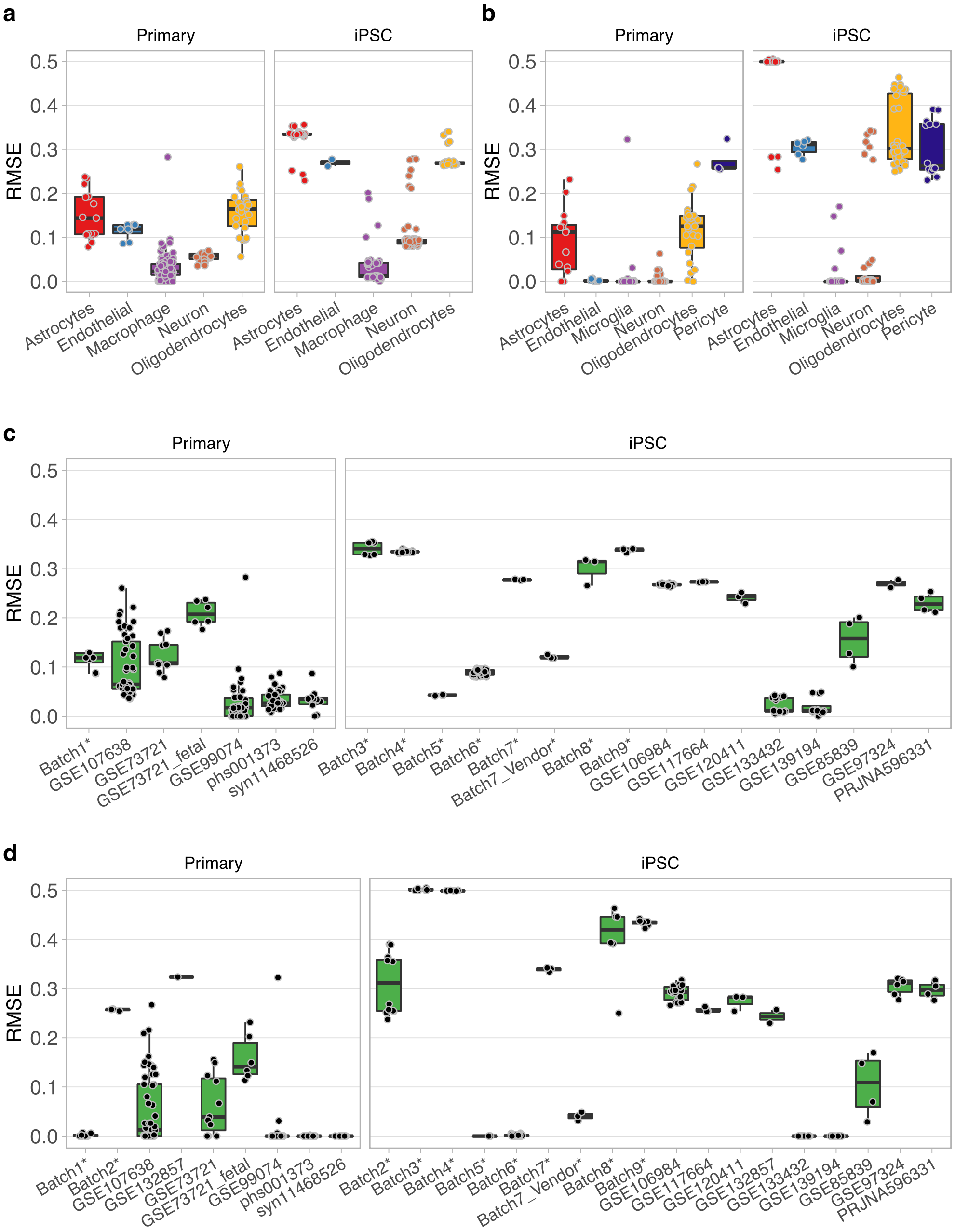


Figure 5. CellMap performance on testing datasets. (**a**,**c**) Independent datasets used for the evaluation of the performance by using the Major9 and (**b**,**d**) CNS6 cell type profiles. (**a**,**b**) RMSE values by cell types with each data point representing one dataset, separated by type of culture into Primary and iPSC datasets. (**c**,**d**) RMSE values by datasets with each data point representing one independent reference bulk sample. Datasets denoted by * refer to in-house characterized cell lines (GSE174379).

increased to nearly 100% by day 7 of the differentiation. Additionally, the cells were resistant to both replating and freeze–thaw cycles as indicated by the unaltered microglia composition (Fig. 6b).

Another in-house bulk RNA-seq dataset was generated to assess batch-to-batch variability of iPSC-derived NGN2 neurons similar to those described by Schmid et al.[24]. The CellMap output indicated low batch-to-batch variability and the cells reaching fully differentiated states by day 21 (Fig. 6c,d). The results showed not only the increasing predicted percentages of neurons as the cultures differentiated, but also the decreases in p-values, indicating that the maturing cells were acquiring transcriptomic profiles more closely resembling those of the primary neurons.

We applied CellMap to samples from the ROSMAP dataset that had matching immunohistochemistry (IHC) and snRNA-seq data available that we used as the ground truth regarding their cell type compositions[18]. What is accepted as the ground truth also affects the apparent success of the prediction. In the case of the ROSMAP samples, comparing the predicted composition to the IHC data resulted in better predictions than using the

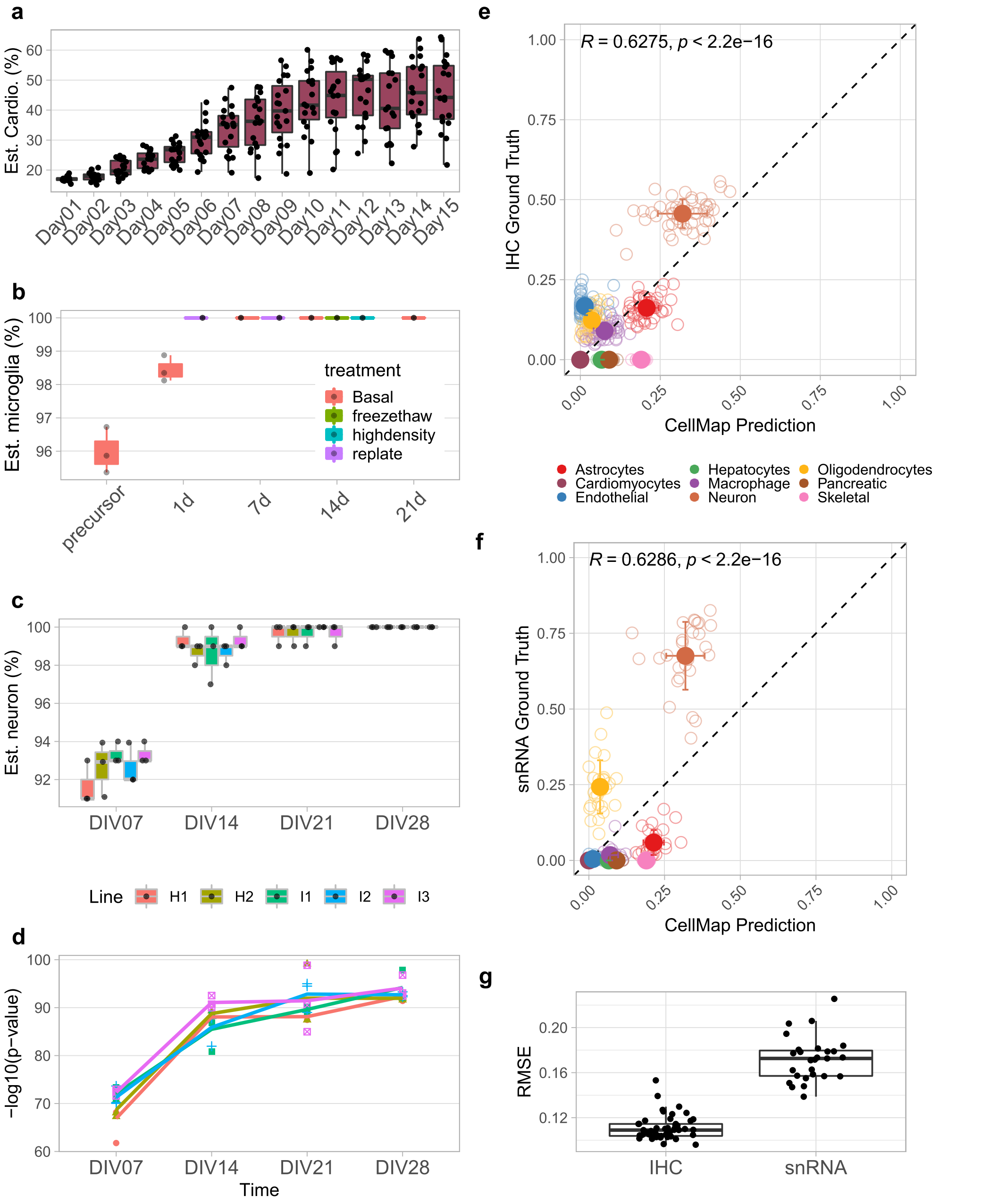


Figure 6. CellMap predictions on bulk RNA-seq data from iPSC-derived cell types and ROSMAP brain samples. (**a**) Time course of iPSC-derived cardiomyocytes from GSE122380. (**b**) Multiple treatments of iPSC-derived Microglia. (**c**,**d**) Comparing five batches of iPSC derived NGN2 neurons. DIV - days in vitro. (**e**,**f**) Sample composition predictions for ROSMAP samples relative to composition measured by IHC and snRNA-seq. Error bars represent the standard deviations across samples for each cell type. (**g**) Composition RMSE values of CellMap predictions across the 9 cell types and all ROSMAP samples compared to the compositions derived from IHC and snRNA-seq data.

cell type information from the matched snRNA-seq data, as there are cell type biases introduced during nuclei isolation (Fig. 6e,f). Similar observations were made previously by Patrick et al.[19] Based on their analysis and the technical variability of snRNA-seq data, we anticipate the IHC proportions to be closer to the true composition. While the overall correlation between the predicted cell type fractions and the IHC and snRNA-seq derived ground truth was practically the same, the composition RMSE values indicated that the CellMap predictions were closer to the IHC values (Fig. 6e–g). CellMap tended to underpredict the neuronal content, while quite accurately predicted the astrocyte components. The oligodendrocyte and endothelial contents were underestimated, seemingly substituted by other cell types. Endothelial cells are especially easily confounded with other cell types. While they present a set of common features, they also possess considerable heterogeneity depending

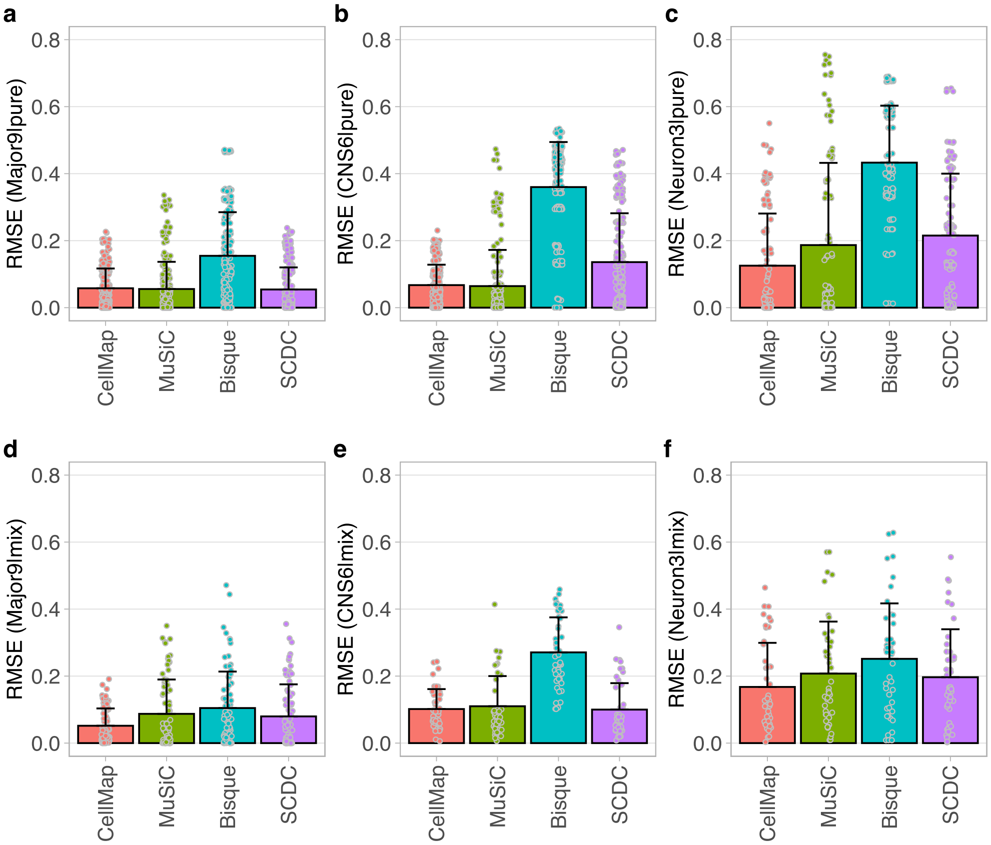


Figure 7. Prediction RMSE values obtained with CellMap, MuSiC, Bisque and SCDC. (**a**–**c**) Comparisons done using the pseudo-bulk pure samples. (**d**–**f**) Comparisons across the pseudo-bulk mixed samples. The tests were performed separately on the training datasets used for generating (**a**,**d**) the Major9, (**b**,**e**) CNS6 and (**c**,**f**) Neuron3 profile sets. Each dot represents one pseudo-bulk sample generated with cells from one of the training datasets.

on their local environment in various organs and tissues[25]. Furthermore, differences in size and RNA content of various cell types can also influence the accuracy of the prediction of cell type proportions[26].

Comparison to other deconvolution tools. The comparisons to the other deconvolution methods were performed on the pure (single cell type) and mixed (multiple cell types) pseudo-bulk samples that were generated as training datasets for deriving the Major9, CNS6 and Neuron3 profile sets. The RMSE values were calculated by comparing the expected and predicted compositions across the individual cell types separately and also for each pseudo-bulk sample across all cell types of a cell type group (Major9, CNS6, Neuron3).

CellMap performed better than Bisque in each of the categories based on RMSE value, while its performance was very similar or slightly better than that of MuSiC and SCDC in predicting the composition of the pseudo-bulk samples (Fig. 7). The range of RMSE values was smallest for CellMap, indicating more consistent predictions across different reference data platforms and cell types. We attribute this robustness to the use of normalization applied to the pseudo-bulk samples as part of the cell type profile generation. In the most difficult task of differentiating between neuronal subtypes (Neuron3 set), CellMap outperformed each of the other three methods (Fig. 7c,f). More detailed comparisons by cell types and by input datasets are provided in the supplementary material separated by input datasets used for the Major9, CNS6 and Neuron3 profile sets (Supplementary Figs. S4, S5, and S6). While its performance is on par with other existing methods, CellMap has the major advantage of flexibility in using reference datasets with non-overlapping cell types and the ability to expand the cell type repertoire with more cell types of interest as reliable and good quality sn/sc RNA-seq datasets become available.

Generating the cell type profiles for CellMap took 23.7 h for the Major9 profiles and 4.25 h and 2.34 h for the CNS6 and Neuron3 profiles, respectively. But this step must be done only once, and the generated profiles can be reused until there are new reference datasets to be incorporated into the profiles. The deconvolution itself was very fast, usually taking less than a minute on a simulated bulk dataset of 220 pure and 80 mixed samples. A deconvolution execution time comparison between CellMap and the other three tools is shown in Supplementary Fig. S7. They were all tested on the same pure and mixed sample sets using 32 processors.

Discussion

We have demonstrated that CellMap can be applied with good accuracy to evaluate cell composition in bulk RNA-seq data. A key feature of CellMap is its ability to be readily re-trained as new sc/sn RNAseq datasets become available and therefore improve the predictions. This information may come "for free" as CellMap is a downstream analysis on RNA-seq data that may have been generated for other purposes (e.g., differential expression or pathway analysis).

Discrepancies between cell type labeling in different sc/sn datasets can result in inaccurate cell type profiles which have a detrimental effect on the deconvolution accuracy[20,21]. Instead of discarding such datasets, a possible solution is to test the consistency of cell type assignments across reference datasets with the help of label transfer algorithms and eliminating only the cells that have conflicting annotations[27,28].

Accounting for the unknown cell type in a bulk dataset is one of the most difficult problems in deconvolution analysis. The undetermined percentage in CellMap's prediction is calculated as the sum of coefficients with insignificant p-values from the overall fitting. When leaving out one cell type profile from a reference set (e.g., leaving out neurons from the CNS6 reference profiles) and then predicting the composition of a mixture that contains the omitted cell type, we generally see the omitted cell type being re-distributed among the remaining cell types, with the most similar ones to the omitted capturing a larger percentage. The undetermined portion generally accounts for 0–5% of the total composition even when the true contribution of the omitted cell type is much higher. Very similar observations were made by Avila Cobos et al. when testing the effect of removing cell types from the reference matrix in multiple deconvolution algorithms[5]. They have found that removing a cell type from the reference matrix resulted in distorted proportions for the other cell types. It is important to note that the CellMap reported p-values indicating the goodness of matches for the different predicted cell types in such cases usually worsen. This is an indication that the analyzed sample might have other cell types not included in the panel of profiles (Supplementary Fig. S8).

Besides the presence of cell types with no defined expression profiles, undetermined cell types in a composition prediction were found when certain cell types had low contributions. To test the accuracy of the deconvolution as a function of cell type abundance in a mixture, pseudo bulk samples with varying compositions were generated from three of the training datasets. In agreement with the findings of Avila Cobos et al. regarding decreased accuracy at low cell type contributions when testing other deconvolution tools, the accuracy of the CellMap prediction tended to be lower when the true contribution of a cell type was below 5%[5]. These low abundance cell types were most often assigned high coefficient p-values and thus were categorized as undetermined (Supplementary Fig. S9).

Another potential source of undetermined cell types in CellMap predictions is the presence of undifferentiated or very immature cells at early timepoints in iPSC-derived cell lines. Two timecourse experiments using iPSC-derived cell lines were selected to illustrate this. In the composition prediction on NGN2 neurons (same dataset as the one used for Fig. 6c,d) at days 7 to 28 of differentiation, CellMap captured some "unintended" cell types at the first timepoint. Afterwards, the cells were already neuronal-like and were identified as neurons. While the CellMap computed p-values were not intended to reflect the extent of cell maturity in general, the decreasing p-values symbolized by increasing number of stars in the bars in this case indicated that the similarity between the profiles of the iPSC-derived NGN2 neurons and those obtained from the mature neurons was increasing with time. The best matches suggesting that neurons with expression profiles closest to the mature neuronal profiles were obtained at day 28 (Supplementary Fig. S9).

In another iPSC-derived pericyte dataset, the CellMap predictions clearly indicated the decreasing contribution of undetermined cell types and the increasing amounts of pericytes from day 0 to day 10 of differentiation. The sample compositions at the latest timepoint were similar to those in published datasets of human primary pericytes. Besides the expected pericytes, each sample had a non-negligeable endothelial component. The close interactions and extensive crosstalk between these two cell types might explain the partial similarity between their expression profiles and, thus, the presence of the endothelial components in the predictions (Supplementary Fig. S9).

CellMap was designed with the characterization of human iPSC-derived cell types in mind. However, it is necessary to acknowledge that iPSC-derived cells may not match primary cells exactly. Even primary cells kept in culture for a relative short time differ from their freshly isolated counterparts. When we compared the transcriptome profile of our in-house generated iPSC-derived microglia to the primary microglia dataset generated by Gosselin and his colleagues, it was not surprising to find out that our iPSC-derived cells were more similar to the cultured primary cells than the freshly isolated ones[29]. It would therefore be unreasonable to expect perfect predictions on iPSC-derived cells when training on the transcriptomic profiles of fully differentiated, mature primary cells.

Despite these shortcomings, deconvolution tools like CellMap are very useful and readily deployed in a cell line development workflow. In combination with microscopic observations of morphological features, they allow us to evaluate the progression of the differentiation in a time-course experiment to ensure correct cell fate. They are excellent quality control methods for verifying batch-to-batch consistency of iPSCs and provide valuable guidance in differentiation protocol optimization.

Data availability

The CellMap R package, including the R code to generate the manuscript figures, is available from https://github.com/interactivereport/CellMap. Pseudo-bulk test datasets with known compositions are provided in the GitHub repository of CellMap https://github.com/interactivereport/CellMap/tree/master/manuscript/pseudo. The in-house RNA-seq data generated from iPSC-derived cell lines has been deposited in the Gene Expression Omnibus

(GEO tracking number GSE174379, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE174379). All other datasets are publicly available and are listed in the supplementary material.

Received: 11 November 2021; Accepted: 10 October 2022
Published online: 17 October 2022

References

1. Chang, C. Y. *et al.* Induced pluripotent stem cell (iPSC)-based neurodegenerative disease models for phenotype recapitulation and drug screening. *Molecules* **25**, 2000 (2020).
2. Engle, S. J. & Puppala, D. Integrating human pluripotent stem cells into drug development. *Cell Stem Cell* **12**, 669–677 (2013).
3. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
4. Knight, A. Animal experiments scrutinised: Systematic reviews demonstrate poor human clinical and toxicological utility. *Altex* **24**, 320–325 (2007).
5. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11**, 5650 (2020).
6. Jin, H. & Liu, Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome Biol.* **22**, 102 (2021).
7. Gaujoux, R. & Seoighe, C. Cell Mix: A comprehensive toolbox for gene expression deconvolution. *Bioinformatics* **29**, 2211–2212 (2013).
8. Radley, A. H. *et al.* Assessment of engineered cells using Cell Net and RNA-seq. *Nat. Protoc.* **12**, 1089–1102 (2017).
9. Cahan, P. *et al.* Cell Net: Network biology applied to stem cell engineering. *Cell* **158**, 903–915 (2014).
10. Newberg, L. A., Chen, X., Kodira, C. D. & Zavodszky, M. I. Computational de novo discovery of distinguishing genes for biological processes and cell types in complex tissues. *PLoS ONE* **13**, e0193067 (2018).
11. Regev, A. *et al.* The human cell atlas. *Elife* **6** (2017).
12. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012).
13. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).
14. Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat. Commun.* **11**, 1971 (2020).
15. Dong, M. *et al.* SCDC: Bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief. Bioinform.* **22**, 416–427 (2021).
16. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
17. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
18. De Jager, P. L. *et al.* A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci. Data* **5**, 180142 (2018).
19. Patrick, E. *et al.* Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLoS Comput. Biol.* **16**, e1008120 (2020).
20. Pasquini, G., Rojo Arias, J. E., Schafer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.* **19**, 961–969 (2021).
21. Shao, X. *et al.* scCATCH: Automatic annotation on cell types of clusters from single-cell RNA sequencing data. *iScience* **23**, 100882 (2020).
22. Sun, J. *et al.* Transplantation of hPSC-derived pericyte-like cells promotes functional recovery in ischemic stroke mice. *Nat. Commun.* **11**, 5196 (2020).
23. Strober, B. J. *et al.* Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).
24. Schmid, B. *et al.* Generation of two gene edited iPSC-lines carrying a DOX-inducible NGN2 expression cassette with and without GFP in the AAVS1 locus. *Stem Cell Res.* **52**, 102240 (2021).
25. Garlanda, C. & Dejana, E. Heterogeneity of endothelial cells. Specific markers. *Arterioscler. Thromb. Vasc. Biol.* **17**, 1193–1202 (1997).
26. Zaitsev, K., Bambouskova, M., Swain, A. & Artyomov, M. N. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nat. Commun.* **10**, 2209 (2019).
27. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
28. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
29. Gosselin, D. *et al.* An environment-dependent transcriptional network specifies human microglia identity. *Science* **356**, eaal3222 (2017).

Acknowledgements

The authors thank Robin Kleiman and Chao Sun from Biogen for their support and expert advice in cell type characterization and in-house NGS data generation. This work was funded in part by the Advanced Regenerative Manufacturing Institute (ARMI) Grant T0042.

Author contributions

This study was conceived of and led by M.I.Z., S.J.E., E.L. and B.Z.. Z.O. designed the algorithm, wrote the computer code, and performed the data analysis with contributions from N.B.T., F.C., X.Z. and K.L.. P.E.C. and S.J.E. lead the production iPSCs and provided scientific insight into the biology of neuronal cell types. P.F.C. and R.C. performed the RNA-seq data generation. M.I.Z., Z.O. and F.C. wrote the paper with feedback from all coauthors.

Competing interests

The authors declare the following competing interests: E.L., P.E.C., P.F.C, R.C., K.L., F.C., S.J.E., B.Z. and M.I.Z. are employees of Biogen and hold stocks from the company. Z.O. and X.Z. are employees of BioInfoRx, Inc. N.B.T. declares no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-22115-1.

Correspondence and requests for materials should be addressed to M.I.Z.