Published online 23 February 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 1 **1**
doi: 10.1093/nargab/lqab011

Single-cell mapper (scMappR): using scRNA-seq to infer the cell-type specificities of differentially expressed genes

Dustin J. Sokolowski[1,2,*], Mariela Faykoo-Martinez[2,3], Lauren Erdman[2,4,5], Huayun Hou[1,2], Cadia Chan[1,2], Helen Zhu[6,7], Melissa M. Holmes[3,8], Anna Goldenberg[2,4,5,9] and Michael D. Wilson[1,2,*]

[1]Department of Molecular Genetics, University of Toronto, Toronto, ON, M5S 1A8, Canada, [2]Genetics and Genome Biology, SickKids Research Institute, Toronto, ON, M5G 0A4, Canada, [3]Department of Cell and Systems Biology, University of Toronto, Toronto, ON, M5S 3G5, Canada, [4]Department of Computer Science, University of Toronto, Toronto, ON, M5S 2E4, Canada, [5]Vector Institute for Artificial Intelligence, MaRS Centre, Toronto, ON, M5G 1M1, Canada, [6]Department of Medical Biophysics, University of Toronto, Toronto, ON, M5G 1L7, Canada, [7]Princess Margaret Cancer Center, University Health Network, Toronto, ON, M5G 2C1, Canada, [8]Department of Psychology, University of Toronto Mississauga, Mississauga, ON, L5L 1C6, Canada and [9]CIFAR, MaRS Centre, Toronto, ON, M5G 1M1, Canada

Received July 24, 2020; Revised December 23, 2020; Editorial Decision February 01, 2021; Accepted February 04, 2021

ABSTRACT

RNA sequencing (RNA-seq) is widely used to identify differentially expressed genes (DEGs) and reveal biological mechanisms underlying complex biological processes. RNA-seq is often performed on heterogeneous samples and the resulting DEGs do not necessarily indicate the cell-types where the differential expression occurred. While single-cell RNA-seq (scRNA-seq) methods solve this problem, technical and cost constraints currently limit its widespread use. Here we present single cell Mapper (scMappR), a method that assigns cell-type specificity scores to DEGs obtained from bulk RNA-seq by leveraging cell-type expression data generated by scRNA-seq and existing deconvolution methods. After evaluating scMappR with simulated RNA-seq data and benchmarking scMappR using RNA-seq data obtained from sorted blood cells, we asked if scMappR could reveal known cell-type specific changes that occur during kidney regeneration. scMappR appropriately assigned DEGs to cell-types involved in kidney regeneration, including a relatively small population of immune cells. While scMappR can work with user-supplied scRNA-seq data, we curated scRNA-seq expression matrices for ∼100 human and mouse tissues to facilitate its stand-alone use with bulk RNA-seq data from these species. Overall, scMappR is a user-friendly R package that complements traditional differential gene expression analysis of bulk RNA-seq data.

INTRODUCTION

RNA-seq is a powerful and widely used technology to measure transcript abundance and structure in biological samples (1). RNA-seq analyses typically compare transcript abundance between conditions by identifying differentially expressed genes (DEGs) (2,3). When RNA-seq of a whole tissue (bulk RNA-seq) is completed, it is often a challenge to determine the extent to which changes in gene expression are due to changes in cell-type proportion (4). This challenge is addressed by single-cell RNA-seq (scRNA-seq) methods that measure gene expression at a single-cell resolution. Despite many advances, technical limitations (e.g., low gene detection per cell and cell dissociation optimization) and cost currently limit the use of scRNA-seq for hard-to-dissociate cell-types and large study designs (5–7). Importantly, several bioinformatics methods that leverage scRNA-seq to learn about cell type proportions (RNA-seq deconvolution) from bulk RNA-seq or leverage bulk RNA-seq to decrease drop-out in scRNA-seq demonstrate the highly complementary nature of these two technologies (8–17).

Single cell RNA-seq experiments readily indicate combinations of genes that are involved in the biological functions altered in an experiment or clinical condition. The value of these data is reflected in the growing number of reposi-

*To whom correspondence should be addressed. Tel: +1 416 813 7654 (Ext 328699); Email: michael.wilson@sickkids.ca
Correspondence may also be addressed to Dustin J. Sokolowski. Tel: +1 416 813 7654 (Ext 328699); Email: dustin.sokolowski@sickkids.ca

tories containing publicly available reprocessed scRNA-seq data, such as PanglaoDB (18), scRNAseqDB (19), SCPortalen (20), Single Cell Expression Atlas (21) and the Human Cell Atlas (22), and conquer (5) that allow for a consistent, tissue-aware reference to the cell-type specificity of individual genes. These initiatives and compiled datasets are valuable resources that can be used to interrogate cell-type specific gene expression and enhance bulk RNA-seq analyses in the absence of a matched scRNA-seq experiment.

RNA-seq deconvolution is a powerful tool that can use scRNA-seq data to infer the relative cell-type proportions of a bulk RNA-seq sample. Estimated cell-type proportions can be directly compared between conditions to identify alterations in cell-type composition (23,24). Bioinformatic tools, such as csSAM (4) and subsequently released Bseq-sc (25), utilize estimated cell-type proportions in bulk RNA-seq data to identify DEGs that were not considered differentially expressed from bulk differential analysis alone (2,3,26). While the *de novo* discovery of cell type specific DEGs is powerful, this analysis requires a large number of samples (e.g., 82 sample were used to identify *de* novo cell-type specific DEGs across three cell-types in Baron *et al.*, 2016) (4,25). Since typical exploratory bulk RNA-seq experiments looking for DEGs do not include as large sample numbers, there is a need for new methods that can leverage the growing number of scRNA-seq reference datasets to interpret typical bulk-RNA-seq experiments.

Here we present a bioinformatic method called single-cell mapper (scMappR). scMappR simultaneously infers which cell-types are driving the expression of a particular DEG and uses that inferred cell-type specificity of DEGs to complete cell-type specific pathway analysis. To do this, we first needed to design scMappR to be a convenient pipeline that connects established bulk RNA-seq DEG analysis workflows to scRNA-seq compendiums. To infer which cell-types are driving the expression of a particular DEG, the scMappR workflow begins by using established deconvolution tools to infer cell-type proportions. With cell-type proportions computed by scMappR and cell-type markers and bulk DEGs imported into scMappR, our method normalizes, scales and integrates these summary statistics. Specifically, scMappR re-weights the fold-change of every DEG by the cell-type specific expression of each gene and the proportions of each cell-type into values we call 'cell-weighted Fold-changes' (cwFold-changes). scMappR then returns a matrix of fold-changes re-calibrated to each cell-type, a list of genes whose differential expression is driven by cell-type specific changes in gene expression, and cell-type specific pathway analysis. We first simulate bulk and cell-type specific RNA-seq samples and show that scMappR both increases the correlation between simulated bulk DEGs and simulated scRNA-seq DEGs while simultaneously assigning bulk DEGs to their correct cell-type. We then demonstrate that scMappR can identify validated cell-type specific gene expression by taking advantage of a reference data set (27) where bulk RNA-seq was performed on cell-sorted samples. Finally, we show that scMappR can identify bonafide differential gene expression changes emanating from a minority cell population present in the mouse kidney during regeneration (14,28). Overall, scMappR is a freely available R package available together with extensive vignettes on CRAN that provides important cell-type specificity to a set of user-provided DEGs.

MATERIALS AND METHODS

Rationale behind measuring cell-type specificity in bulk RNA-seq data

We developed a statistic called cell-weighted fold-changes (cwFold-change). cwFold-changes incorporate cell-type specificity and cell-type proportion to infer which cell-types are likely driving a DEG identified in bulk RNA-seq data (Supplementary Figure S1). The cwFold-change metric accounts for expression differences driven by cell-type proportion differences in the bulk samples (Supplementary Figure S1F). Our package, scMappR, measures cwFold-changes from imported bulk DEGs and identifies which bulk DEGs may be cell-type specific before sorting them to their appropriate cell-type. scMappR then re-orders DEGs by their cwFold-change to complete cell-type specific pathway analysis.

Defining the statistics scMappR uses to re-weight bulk differentially expressed genes

We use cell-type specific gene expression, cell-type proportion of a bulk sample, and the fold-change of a gene's expression between conditions to re-weight a bulk DEG in a cell-type specific manner. We define cell-type specificity as the weighted sum of cell-type specific gene expression (Equation 1) and cell-type proportion (Equation 2) because in bulk RNA-seq, all RNA from a sample is pooled together. We calculate the relative amount of RNA originating from each cell-type as the product of cell-type specificity and cell-type proportion.

$$P_c = \frac{N_c}{\sum_{c=1}^{C} N_c} \tag{1}$$

where P represents proportion, c represents cell-type, N_c represents the number of cells of type c.

$$S_{c,g} = \frac{\text{mean}\left(E_{c,g}\right)}{\text{mean}\left(E_{t\neq c,g}\right)} \tag{2}$$

where S represents specificity, E represents expression, and g represents gene.

$$K_{c,g} = P_c * S_{c,g} \tag{3}$$

where K represents cell-type contribution.

The fold-change of a DEG is the ratio of means in gene expression between conditions (Equation 4).

$$F_g = \frac{\text{mean}\left(E_{g,y=1}\right)}{\text{mean}\left(E_{g,y=0}\right)} \tag{4}$$

F represents the fold-change of a DEG.

Normalizing for the dependence between cell-type specificity and cell-type proportion

We use RNA-seq deconvolution to estimate cell-type proportions in scMappR; however, RNA-seq deconvolution

requires cell-type specificity as an input to measure cell-type proportions in the bulk sample (13,15,16,29–31). In scMappR, we developed an RNA-seq deconvolution normalization step to allow the expression of each DEG to be independent from inferred cell-type proportions. We re-calculate cell-type proportions for each DEG after iteratively removing the DEG from the bulk normalized count matrix and signature matrix. A signature matrix is defined as a gene-by-cell-type matrix containing the fold-change difference between a given cell-type and all other cell-types (Equation 2). This normalization step yields an estimated cell-type proportion for every DEG, where the proportions are independent of that DEG's expression. We could then assign cell-type specificity to the fold-change of a DEG with the knowledge that cell-type expression and cell-type proportion are independent.

Correcting for differentially expressed genes driven by changes in cell-type proportion

scMappR accounts for cell-type composition because a gene may be detected as differentially expressed due to differences in cell-type proportions alone (Supplementary Figure S1F). We account for differences in cell-type proportion in scMappR by adding a scaling factor to the cell-type partitioning of DEGs. Specifically, for each DEG, we calculate the average cell-type proportions in each condition, and we scale the DEG by the reciprocal ratio between the two conditions. Specifically, if the DEG is biased to one condition (e.g., condition 2), then we scale the DEG by the relative difference in cell-type proportions between the conditions (Equations 5–7).

$$\frac{P_{c,g}|y\ =\ 0}{P_{c,g}|y\ =\ 1} \tag{5}$$

$$\widetilde{F_{c,g}} = \frac{\text{mean}\left(E_{g,y=1}\right)}{\text{mean}\left(E_{g,y=0}\right)} * \frac{P_{c,g}|y\ =\ 0}{P_{c,g}|y\ =\ 1} \tag{6}$$

$\widetilde{F_{c,g}}$ represents the cell-type proportion scaled fold-change of a DEG.

$$\widetilde{F_{c,g}} = \frac{\text{mean}\left(E_{g,y=1}\right)}{P_{c,g}|y=1} * \frac{P_{c,g}|y\ =\ 0}{\text{mean}\left(E_{g,y\ =\ 0}\right)} \tag{7}$$

Re-weighing differentially expressed genes by cell-type specific information by generating cell-weighted fold-changes

After the normalization and scaling steps are complete, we multiply fold-change, cell-type specificity and cell-type proportions to partition the cell-type specificity of the DEGs originating within the users' bulk sample. We call this re-weighted fold-change of a DEG a cell-weighted fold-change (cwFold-change, Equation 8, Figure 1).

$$cwF_{c,g} = \widetilde{F_{c,g}} * \widetilde{K_{c,g}} \tag{8}$$

$cwF_{c,g}$ represents the cwFold-change.

The following criteria are met when calculating cwFold-changes. If the cell-type specificity of a gene for a cell-type is <1, then the cwFold-change for that cell-type decreases in its contribution to the DEG. If the cell-type specificity is >1, then the cwFold-change for that cell-type increases in its contribution to the DEG (Supplementary Figure S1 D and E). If the proportion of a cell-type in the bulk sample reaches 0, then the DEG did not originate from that cell-type. If the proportion of a cell-type in the bulk sample reaches 1, then the DEG must originate entirely from that cell-type.

Estimating cell-type proportions in scMappR

scMappR requires an RNA-seq deconvolution method that has a signature matrix as input in order to perform the normalization step that allows for cell-type proportions to be independent of the imported signature matrix. Computational run-time is also a consideration in the cell-type proportion and normalization steps within scMappR (e.g., if there are 3500 DEGs, RNA-seq deconvolution is repeated 3501 times). We currently have three RNA-seq deconvolution tools incorporated into scMappR. These deconvolution tools are DeconRNA-seq (16), whole genome correlation network analysis (WGCNA) (30), and the digital cell quantifier in ComICS DCQ (29) through the original and ADAPTS (32) R package. DeconRNA-seq (16) is a quadratic programming-based deconvolution tool that minimizes cell-type proportions based on the signature matrix and normalized RNA-seq counts. DCQ (29) is a regression-based deconvolution algorithm that relies on elastic net regularization. WGCNA (30) uses the 'proportionsInAdmixture' function and integrates correlation and linear least-squares regression. In the present study, we used DeconRNA-seq (16) (Equation 9) as the primary scMappR RNA-seq deconvolution tool.

$$\min_{\text{p}} \left(||PS - X||^2\right), s.t. \{ \begin{smallmatrix} \sum_c P_c = 1 \\ P_c \geq -, \forall c \end{smallmatrix} \tag{9}$$

X represents normalized bulk gene expression.

Identifying cell-type specificity in cwFold-changes

cwFold-changes are evaluated at the level of the gene and at the level of the cell-type. Evaluating cwFold-changes at the level of the gene identifies which cell-types may be driving a DEG. We first normalized the cwFold-change for each gene so that it sums to 1, allowing for a more interpretable measure of the contribution each DEG has on a cell-type specific fold change. The distributions of normalized cwFold-changes were then tested for a normal distribution with a Shapiro test (33) within each gene (across cell-types). Cell-type outliers were measured according to whether the distribution normalized cwFold-change was parametric (Equation 10) or nonparametric (Equation 11). We ensure that the cell-types driving these DEGs have a normalized cwFold-change greater than the cell-type proportion and a uniform distribution (Equation 12). cwFold-changes for each cell-type must be less than the cell-type proportion for that cell-type. This filter prevents genes from being assigned to a cell-type because the cell-type is very common in the bulk sample. In this filtration step, we allow the user to import cell-type proportions (P_c, Equation 12) from any source (e.g., xCell, CIBERSORT, cell population mapping, Mu-

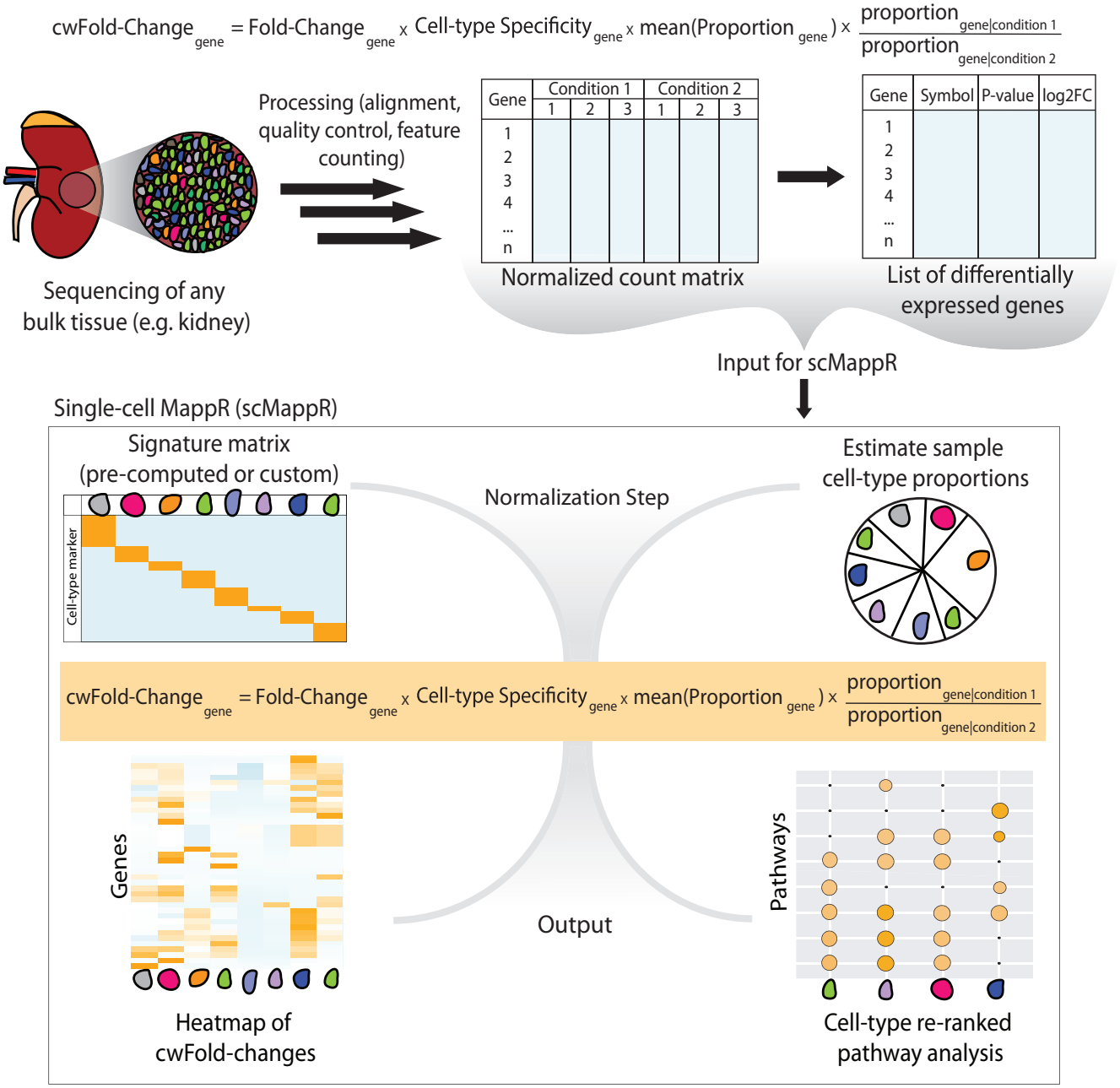

Figure 1. Schematic of the data required to run scMappR and the primary functionalities that scMappR provides. scMappR requires input RNA-seq count data, a list of differentially expressed genes, and a signature matrix (provided by the user or scMappR). For each gene, scMappR then makes cell-type expression independent of estimated cell-type proportions. scMappR then integrates cell-type expression, cell-type proportion, and the ratio of cell-type proportions between biological conditions to generate cell-weighted Fold-changes (cwFold-changes). These cwFold-changes are then visualized (bottom left) and re-ranked before scMappR computes and plots cell-type specific pathway analyses (bottom right).

SiC) (13,15,34,35).

$$\text{upperBound}_g = \text{mean}\left(\widetilde{cwF_g}\right) * 3\, sd\left(\widetilde{cwF_g}\right) \quad (10)$$

$\widetilde{cwF}$ represents the normalized cwFold-change

$$\text{upperBound}_g = \text{median}\left(\widetilde{cwF_g}\right) * 3\, \text{mad}\left(\widetilde{cwF_g}\right) \quad (11)$$

$$\text{Specific}_{c|g} = cw\text{FoldChange}_{c|g}\ s.t. \begin{cases} \widetilde{cwF_{c,g}} > \text{upperBound}_g \\ \widetilde{cwF_{c,g}} > P_c \\ \widetilde{cwF_{c,g}} > \frac{1}{N} \end{cases} \quad (12)$$

cwFold-changes are evaluated at the level of the cell-type by re-ordering each DEG by their cwFold-change in each cell-type. The rank order of bulk DEGs and cwFold-changes for each cell-type are measured with a Spearman's correlation.

Interpretation of cwFold-changes at the level of the cell-type through cell-type specific pathway enrichment

Cell-weighted fold-changes are computed for every DEG in each cell-type. cwFold-changes and endogenous cell-type specificity are then plotted with the Pheatmap R package (36). For every cell-type, each DEG on the gene list is re-ranked by their cwFold-change. We use a cwFold-change cutoff of a gene in a cell-type as 10^{-10} to determine if a gene is within a cell-type. Genes that do not meet this cutoff are discarded from the pathway analysis for that cell-type. DEGs that change in expression to a similar degree may be under a common biological regulator (30,37). Having the same list of DEGs but in a different order of significance can have a profound impact on which pathways are enriched. Pathway analysis is subsequently completed with g:ProfileR package (38). By default, scMappR uses the following example command: 'gprofiler(genes, species, ordered = TRUE, src_filter = c('GO:BP', 'REAC', 'KEGG'), custom_bg = genes_in_bulk, correction_method = 'fdr')' (38,39) (Figure 1). We report precision as the g:Profiler summary statistic which g:Profiler defined as the proportion of the DEGs that are present in the gene set (38).

Implementation

We present the R package scMappR to infer cell-type specificity in bulk DEGs. scMappR can be installed from CRAN (Supplementary File S1: page 1) and we provide a full workflow for a researcher to convert bulk RNA-seq data into cwFold-changes in Supplementary File S1.

In the context of bulk RNA-seq, scMappR expects that the normalized data are already pre-processed to account for any batches, artifacts and read depth. Transcript-based approaches to counting bulk RNA-seq should be mapped to their genes as scRNA-seq data are counted at gene level. These bulk data can be the same input used for most RNA-seq analyses pipelines including data visualization such as PCA, RNA-seq deconvolution tools or correlational approaches such as Whole Gene Network Correlation Analysis (15,16,30,40). We used counts-per-million in Supplementary File S1 but any pre-processing method is acceptable. scMappR further expects that the list of imported differentially expressed genes were measured accounting for potential appropriate factors (e.g. batch). Users should account for the uncertainty of their bulk DEGs and cell-type markers by applying relevant multiple-test corrected *P*-value and fold-change cutoffs to their DEGs and cell-type markers. We use an FDR adjusted *P*-value cutoff of 0.05 and an absolute fold-change cut-off of 1.5 in Supplementary File S1.

Similarly, users should evaluate if the cell-type proportions estimated in scMappR are reflective of their bulk samples (Supplementary File S1: page 2). Evaluating uncertainty in their cell-type proportion and applying the appropriate *P*-value cutoffs to imported bulk DEGs and cell-type markers within the imported signature matrix is an important step prior to scMappR because the cwFold-change metric does not inherently contain uncertainty.

With RNA-seq data normalized (Supplementary File S1: pages 2-3), bulk DEGs are calculated (Supplementary File S1) and RNA-seq deconvolution methods are compared (Supplementary File S1: page 3). The user needs to index which samples are associated with 'upregulated' and 'downregulated' DEGs before calculating cwFold-changes and cell-type specific pathway enrichment (Supplementary File S1: pages 3-4).

Finally, the 'cwFoldChange_evaluate' function is used to investigate the cell-type specificity of their cell-types and sort DEGs into different cell-types (Supplementary File S1: pages 4, 6). While we import cell-type proportions calculated within scMappR, this matrix of cell-type proportions for each sample may be imported from any tool (e.g. xCell, MuSiC, cell population mapping, CIBERSORT) by the user (13,15,34,35).

Generating simulated RNA-seq data

Simulated reads were generated using the Polyester R package (41). Overall, we simulate five cell-types, two conditions and six simulated replicates summing to a total of 60 simulated samples (Figure 2A). Specifically, each simulated read was 100 base pairs and each simulated RNA-seq experiment had 20× coverage. Simulated reads were based on 20 000 genes from the Gencode human hg38 genome (release 31) (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_31/gencode.v31.transcripts.fa.gz). We simulated five different cell-types for each simulated sample. In our simulations, 30% of simulated genes contained a difference in fold-change between one cell-type and all others; this represented cell-type specificity. The degree of cell-type specificity (i.e. the fold-change between cell-types) is a variable that we adjusted in the different iterations of the simulation. For each cell-type, we simulated two conditions. For each pair of simulated conditions, we set the estimated fold change to three. We ensure that the fold-change between conditions does not influence the cell-type specificity of a gene by increasing the 'upregulated' condition and decreasing the 'downregulated' condition. This way, the genes that we simulated to have a difference in means across conditions do not also have a difference in means across cell-types unless specified. We set 5% of simulated genes to be differentially expressed between conditions in one cell-type, and another 5% of simulated genes to be differentially expressed between conditions in all cell-types. Finally, we simulated six replicates for each cell-type to allow for enough variation to generate DEGs.

Our simulated bulk RNA-seq samples are a weighted sum of the simulated cell-type specific samples. Each of the simulated replicates have counts for all five cell-types. The expression of each bulk gene is described in Equation (13). Each time we simulate bulk RNA-seq data we designate the proportion of each cell-type, thus controlling cell-type composition.

$$\text{BulkCount}_{\text{g}} = \sum_{c=1}^{C} \text{Count}_{g,c} \tag{13}$$

Simulated batch effects in bulk RNA-seq data

We generated batch effects using two different strategies because batch effects may arise from many different reasons. In both strategies, we simulated six samples, half from each

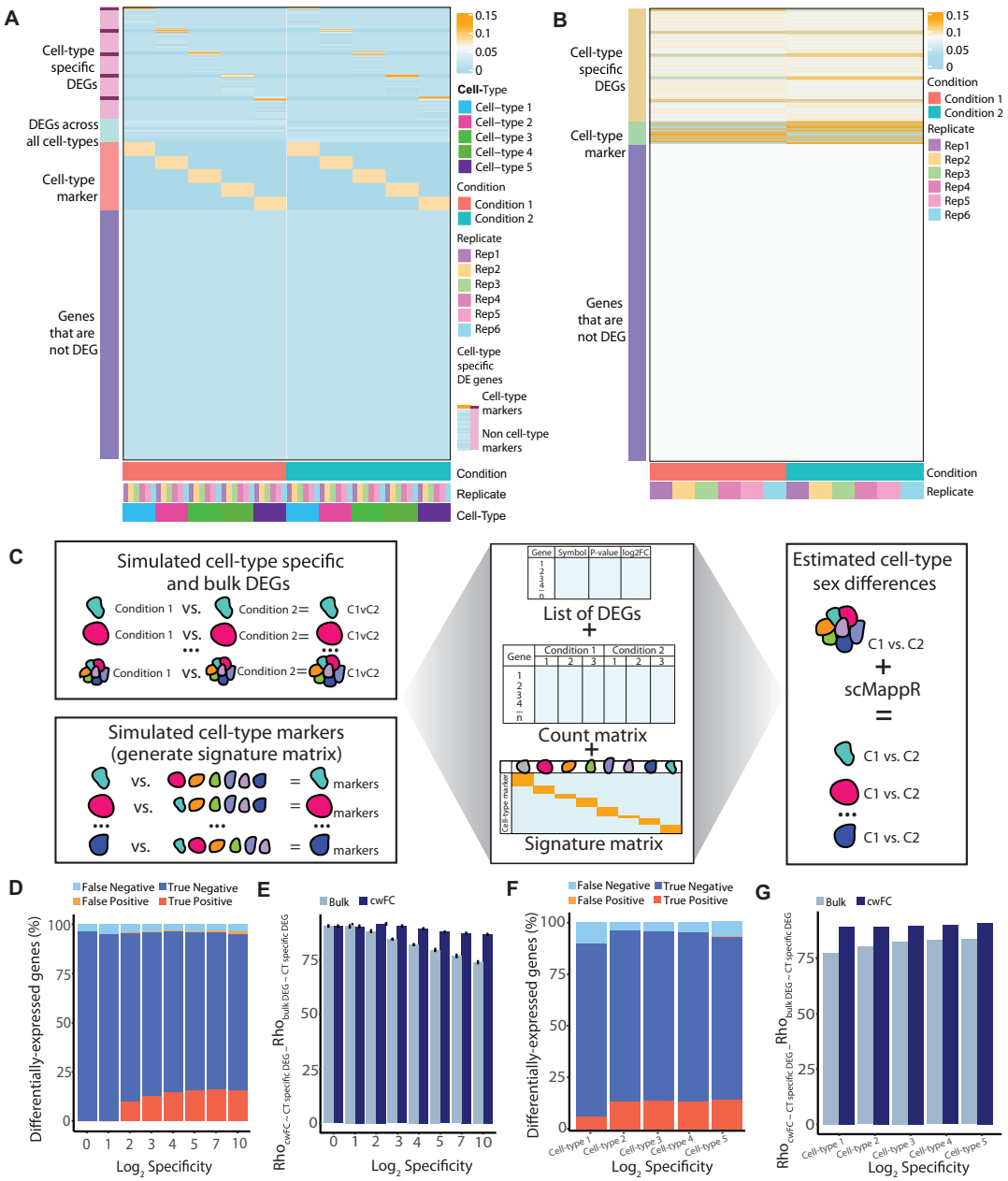

Figure 2. Evaluation of scMappR performance using simulated RNA-seq data. (**A**) Gene-normalized heatmap of the mean expression of every simulated gene within each simulated sample, condition, and cell-type within simulated cell-type specific RNA-seq data. Each row is a simulated gene, and each column is a simulated cell-type within a simulated sample. This matrix is directly imported into the 'simulate_experiment_countmat' function in the polyester R package to generate cell-type specific RNA-seq fasta files. The bar on the left of the heatmap designates the gene expression profile of our simulated genes. The bars on the bottom of the heatmap designate the simulated cell-type, condition, and replicate of each simulated sample. The legend on the right assigns each condition, cell-type, and replicate (**B**) Gene-normalized heatmap of the simulated bulk RNA-seq samples based on (A). This matrix is directly imported into the 'simulate_experiment_countmat' function in the polyester R package to generate bulk RNA-seq fasta files. The simulated expression of each gene is the mean gene expression of each cell-type weighted by the cell-type proportion that we designate in each iteration of our simulation. Therefore, each row is a simulated bulk gene that matches the row in (A). The bar on the left of the heatmap designates the bulk gene expression profile of our simulated genes. The bars on the bottom designate condition and replicate (corresponding with (A)). The legend on the right assigns each condition, cell-type, and replicate. (**C**) Schematic of evaluating scMappR with simulated data. We measure cell-type specific DEGs and bulk DEGs across conditions. We also measure cell-type specificity by calculating the DEGs between one cell-type and all of the other cell-types. These cell-type specific DEGs become our signature matrix. We then apply scMappR to the simulated bulk RNA-seq, DEGs, and signature matrix before evaluating our predicted cell-type specificity against the simulated cell-type specific DEGs. (**D**) Barplot of the proportion of true/false positives and negatives at different degrees of cell-type specificity. Cell-type proportions for all five cell-types are fixed at 20%. True positive is red, True negative is blue, false positive is light orange, false negative is light blue. (**E**) Average improvement that cell-weighted fold-changes (cwFold-changes) have on cell-type specificity as cell-type specificity increases is measured with a bar chart. Dark bars are the correlation cwFold-changes with cell-type specific fold-changes. Light = bars are the correlation between cell-types (left) and a boxplot of the correlations across cell-types (right). (**F**) Barplot of the proportion of true/false positives and negatives in cell-types with different cell-type proportions. Cell-type specificity is set to a fold-change of 32 between cell-type markers. (**G**) Average improvement that cell-weighted fold-changes (cwFold-changes) have on cell-type specificity for every cell-type is measured with a bar chart.

of the two simulated conditions. The first strategy represents simulated reads from two different sequencing platforms, where the first batch had a sequencing bias of 'rnaf' and an error model of 'illumina4' and the second batch had a sequencing bias of 'cdnaf' and an error model of 'illumina5' using the 'simulate_experiment_countmat' function in Polyester (41). The second strategy represents a batch effect that had a larger impact on gene expression than the condition itself. In this second strategy, we randomly selected 20% of all genes to have a fold-change difference of four (greater than the difference in condition) in either batch 1 or batch 2. In our tests of the influence of batches on scMappR, we fixed the cell-type specificity between the cell-types to have a fold-change of 32, and every one of the five cell-types contained a different proportion of the bulk sample (cell-type 1 = 1/15, cell-type 2 = 2/15, . . . , cell-type 5 = 5/15).

Generating cell-type specific differential expression in simulated reads

We measure cell-type specific expression as the fold-change between one cell-type versus all others. We scaled the fold-change between cell-types as $\log_2$(fold-changes). If $s = 0$, there is no cell-type specificity. If $s < 2$, cell-type specificity has a smaller fold-change than the effect of conditions, which we are defining as our null process. If $s > 2$, cell-type specificity has a larger effect than the effect of the condition. We considered the effect of cell-type identity to be greater than the effect of the condition as the true generative process. When our simulations tested the influence of cell-type composition on scMappR, we fixed simulated cell-type specific genes at a difference in fold-change of 32 ($s = 5$). We considered a bulk sample consisting of all five cell-types to be considered the true generative process. The parameters of our simulated RNA-seq data are summarized in Supplementary Tables S1 and S2.

Getting output of simulated differential analysis and changes in parameters

Simulated RNA-seq reads were aligned to the hg38 genome with the STAR aligner (42) with default paired-end sequencing parameters before being filtered for ENCODE blacklist regions with bedtools (43). Simulated reads were then counted using the featureCounts (44) tool with parameters '–s 1 –Q 1 –t 'exon'. Counted reads were RPKM normalized using edgeR (3). When bulk RNA-seq were simulated in two different batches (see Supplementary Methods), then simulated batches were corrected for use of the Combat-seq within the sva R package (45,46). Differential expression within each cell-type and bulk sample was done to measure DEGs across conditions using the Wald's test (2). When there were batch effects, differentially expressed genes were measured with a likelihood ratio test with 'batch' as a reduced variable in DESeq2 (2). Differential expression between each cell-type and all other cell-types was completed to identify cell-type markers. These cell-type markers were made into a signature matrix using the 'generes_to_heatmap' function in scMappR. We then generated cwFold-changes of bulk DEGs with the generated signature matrix. The process of simulating, aligning and counting RNA-seq data before measuring cell-type DEGs, cell-type markers, bulk DEGs, and cwFold-changes was repeated in each iteration of scMappR's evaluation. We altered the cell-type specificity in the simulated cell-type specific RNA-seq and cell-type proportions in the simulated bulk RNA-seq in each iteration of scMappR's evaluation (Supplementary Tables S1 and S2).

Simulation evaluation parameters

We first define true and false positives and negatives when we evaluate if bulk DEGs are getting assigned to cell-types correctly. Every gene assigned to a given cell-type with the 'cwFold-change_evaluate' function was considered positive for that cell-type while a gene that did not map to that cell-type was considered to be a negative. A true positive is a cwFold-change mapping to a cell-type where the gene is differentially expressed in the cell-type specific RNA-seq. A false positive is a cwFold-change mapping to a cell-type where the gene is not differentially expressed. A true negative is if the bulk DEG does not map to the cell-type and the gene is not differentially expressed within that cell-type, or if the gene is equally differentially expressed in all cell-types. A false negative is if the bulk DEG does not map to the cell-type and the gene is differentially expressed in that cell-type (additional detail in Supplementary Tables S3 and S4).

We then define how we measured whether the distribution of DEGs re-ordered by their cwFold-changes better reflected true simulated cell-type specific DEGs. We correlated simulated bulk DEGs to simulated cell-type specific DEGs. Simultaneously, we correlated simulated cwFold-changes to the same simulated cell-type specific DEGs. We then tested if scMappR improved cell-type specificity by asking if cwFold-changes improved the correlation between bulk and cell-type specific DEGs across cell-types. We test the improvement in correlation between cwFold-changes to cell-type specific DEGs and bulk DEGs to cell-type specific DEGs with a one-tailed paired Student's *t*-test paired by cell-type. This test is one-tailed because we are exclusively testing for the improvement in cell-type specificity.

Generation of cell-type signature matrices from publicly available scRNA-seq

Consistently reprocessed scRNA-seq samples were obtained from bulk data in the PanglaoDB (18) project (https://panglaodb.se/samples.html). Briefly, PanglaoDB (18) automatically downloads mouse and human scRNA-seq data before aligning and processing these data in a manner specific to their sequencing platform (Drop-seq, 10X Genomics, and Start-seq) (47,48). The scMappR package provides the bioinformatic pipeline to convert any scRNA-seq count dataset into a signature matrix with named cell-types within scMappR's 'process_dgTMatrix_lists' function. We use the fold-change (non-log) output in the 'FindMarkers' function within Seurat V3 (49,50) to populate these signature matrices. All normalization, clustering, cell-type marker identification and cell-type labelling steps detailed below describe the 'process_dgTMatrix_lists' function, and

how it was applied to the scRNA-seq data stored in the PanglaoDB database (18). To generate signature matrices from scRNA-seq count data, we removed cells with abnormally high mitochondrial content (greater than two standard deviations above the mean in that given sample) (51). Then, normalization, clustering, scaling and integration of technical replicates were completed using Seurat V3 with the integration anchors feature (49,50). Cell-type markers are identified using the 'FindMarkers' function in Seurat v3 (default parameters) (49,50). This function completes differential expression (Wilcoxon's test as default) between each cell-type and all the other cell-types. Signature matrices are populated with the rank ($-\log_{10}$(*P*-value)) to measure enrichment of a cell-type or fold-change output to calculate cwFold-changes. We further use these cell-type markers to define cell-types. Cell-types were identified by extracting the (at maximum) top 30 cell-type markers and converting each gene symbol to human or mouse symbols when necessary using Ensembl BioMart (52).

Our automated cell-type identification pipeline is based on two gene set enrichment methods, namely the Fisher's exact test of cell-type markers and Gene Set Variation Analysis (GSVA) of the average expression of each gene per cell-type (53–55), against two cell-type marker databases: CellMarker and PanglaoDB (17,18). The CellMarker database manually curated cell-type markers using a literature search of over 100 000 papers and is updated four times per year (17). The PanglaoDB database was generated with a combination of manual curation, co-expression of putative cell-type markers, and community submission (18). scMappR automatically labeled cell-types by appending the cell-type label of the most highly enriched cell-type from the CellMarker database to the most highly enriched cell-type using the PanglaoDB database using a two-tailed Fisher's exact test (17,18,55). Cell-types that do not contain significant enrichment of either database with the Fisher's exact test (55) were labelled unknown, however all cell-types (including unknown) have predicted labels from the GSVA method (53) stored as an output file. Once cell-types were labeled, signature matrices based on rank and fold-change were generated. scMappR reprocesses user-provided scRNA-seq count data with the same pipeline. We aggregated all the cell-types and cell-type markers into a gene-set database. Each gene-set is designated with the following notation: 'SRA ID: tissue: cell-type'. All the cell-type markers within each gene list are consistently processed. This gene-set database can be used for gene-set enrichment using a Fisher's exact test (55) within scMappR and the gene-set database can be downloaded for other gene-set enrichment analysis tools (39).

The bioinformatic pipeline used to process scRNA-seq from count matrix data is part of the scMappR R package. Users can optionally provide their own scRNA-seq count matrix, which is converted into a Seurat (50) object that is then processed and converted into a signature matrix using the same methods described above. Users can additionally choose to save intermediary files generated by scMappR to process count matrices into a signature matrix. Specifically, scMappR saves the Seurat object, all cell-type markers, and all possible cell-type labels from both CellMarker and Panglao (using GSVA and the Fisher's exact test) (17,18,49,50,53–55). Finally, the vignette stored in CRAN provides the functions required to convert a Seurat object into a signature matrix. Together, this pipeline is scMappR's 'process_dgTMatrix_lists' function. It can be used as a consistent scRNA-seq processing pipeline from a count matrix of raw scRNA-seq data.

Processing RNA-seq data from Monaco *et al.*, 2019

All fastq files from the peripheral blood mononuclear cells (PBMC) dataset and 29 fluorescence activated cell sorted (FACS) immune cell-types were obtained from GSE107011 (27) using sratoolkit (56). Samples were aligned to the hg38 genome with the STAR aligner (42) using default parameters for paired-end sequencing and filtered for blacklist regions. Reads were assigned to genes using featureCounts (version 1.5.3) with parameters '-s 1 -Q 255 -t exon -O'. Gene models were obtained from GENCODE v33. Reads per kilobase per million (RPKM) were then calculated for each gene using edgeR and principal component analysis (PCA) was performed (3). Sex differences ($N = 9$ female, 4 male) were measured across the bulk PBMC dataset. Sex differences were also measured in the experiments where RNA-seq was completed after cell-sorting each immune subtype ($N = 2$ female, 2 male). In both cases, differential expression was completed using DESeq2 (Wald's test; FDR adjusted *P*-value < 0.05 and absolute fold-change > 1.5) (2). Cell-type markers were then computed by measuring differential expression of genes in each cell-type against all others (Wald's test; adjusted *P*-value < 0.05 and absolute fold-change > 2). The 'ggplot2' and 'ggfortify' packages were used to generate all plots (40,57).

Processing RNA-seq data from Valle Duraes *et al.*, 2020

All fastq files related to RNA-seq on the bulk kidney were downloaded from ArrayExpress (E-MTAB-7957) using 'wget'. All fastq files related to bulk RNA-seq of T-cells and T-regulatory cells were also downloaded from ArrayExpress (E-MTAB-7961) using 'wget'. These RNA-seq bulk kidney samples were aligned to the mm10 genome with the STAR aligner (42) using default parameters for paired-end sequencing and filtered for blacklist regions. Reads were assigned to genes using featureCounts (version 1.5.3) with parameters '-s 1 -Q 255 -t exon -O'. Gene models were obtained from GENCODE M11. Samples were then RPKM and PCA was performed (40,57). Differential expression analysis between condition (naïve vs. Regeneration, and naïve versus Fibrosis) was performed using DESeq2 (Wald's test; FDR adjusted *P*-value < 0.05 and absolute fold-change > 1.5) (2).

Cell-type marker enrichment from any imported gene list

When users import a generic list of genes and a tissue of interest or a signature matrix, scMappR plots the inputted signature matrix and the signature matrix overlapping with the gene list using the Pheatmap R package (36). scMappR then tests the enrichment of cell-type markers that overlap the user's list with a Fisher's exact test (odds ratio > 0, FDR adjusted *P*-value < 0.05) (55) while using all cell-type markers as a statistical background.

R Package: scMappR

The scMappR R package allows users to interrogate how different cell-types are driving DEGs within a bulk sample. scMappR contains the bioinformatic pipeline needed to process scRNA-seq data from a count matrix to formats compatible with scMappR. The rationale for creating this pipeline was to ensure scMappR could be widely used downstream of existing bulk RNA-seq DEG analyses. scMappR is currently stored on CRAN (https://cran.r-project.org/web/packages/scMappR/index.html). A stable release of the re-processed scRNA-seq data is stored on Zenodo (https://zenodo.org/record/4278129#.X87JuGhKg2w). Later releases of re-processed scRNA-seq data that are updated with new datasets are stored in a separate GitHub repository (https://github.com/wilsonlabgroup/scMappR_Data).

RESULTS

Summary of the scMappR R package and functionality

The primary function of scMappR is to integrate cell-type expression and cell-type proportions to calculate a metric that we call cell-weighted fold-changes (cwFold-changes). cwFold-changes can be evaluated at the level of the gene to determine which bulk DEGs are likely originating from expression in a specific cell-type. cwFold-changes can also be evaluated at the level of the cell-type to infer the distribution of bulk DEGs within each cell-type. Grouping genes that may be under cell-type specific regulatory control has the potential to improve functional enrichment analysis using tools such as GSEA and g:Profiler (37,39).

We calculate and evaluate our cwFold-change metric with the 'scMappR_and_pathway_analysis' function (Figure 1). Users input a list of DEGs, normalized counts and a signature matrix into this function. scMappR then re-weights bulk DEGs by cell-type specific expression from the signature matrix, cell-type proportions from RNA-seq deconvolution (16) and the ratio of cell-type proportions between the two conditions to account for changes in cell-type proportion (Figure 1) (see 'Materials and Methods' for details). Genes containing cell-type specific differential expression are then sorted into the cell-type driving the DEG with the 'cwFoldChange_evaluate' function (see 'Materials and Methods' for details). The 'cwFoldChange_evaluate' function also measures the difference in distribution between bulk DEGs are cwFold-changes in each cell-type. Finally, we re-order DEGs by their cwFold-changes before completing pathway enrichment (38,39). To better understand which pathways are most impacted by cell-type specificity, we re-order DEGs by their difference in fold-change before and after scMappR is applied before completing pathway enrichment.

We designed the scMappR package so that users with DEG list obtained from bulk RNA-seq experiments can quickly classify their list of DEGs based on the likely cell type(s) of origin. To make this process simple we designed scMappR so that the users do not need to obtain and process their own scRNA-seq data. We incorporated well-established scRNA-seq processing and cell-type identification packages (17,18,53,54) in databases into a single pipeline into scMappR (see 'Materials and Methods' for details). We applied this pipeline to uniformly process the PanglaoDB (18) database, where we converted over 1000 re-processed human and mouse scRNA-seq datasets into 331 signature matrices across over 100 tissues (Supplementary Table S5). Users can access these signature matrices with the 'get_signature_matrices' function in scMappR. These signature matrices are directly compatible with scMappR to generate cwFold-changes. Users may also use the scMappR R package to re-process their own scRNA-seq data with the same pipeline where we processed the 331 signature matrices by applying the 'process_dgTMatrix_lists' function to their scRNA-seq data. This function provides more flexibility than using Seurat and the Wilcoxon test exclusively such as choosing different parameters in scRNA-seq normalization (e.g. Seurat or scTransform) (49,50,58) and cell-type marker techniques (e.g. Wilcoxon test, MAST, bimod) (59,60).

Testing scMappR using simulated RNA-seq data

We used a series of simulations (41) to evaluate whether scMappR properly assigns bulk DEGs to their correct cell-type and if scMappR makes the distribution of bulk DEGs better reflect the distribution of cell-type specific DEGs (see 'Materials and Methods' section). These simulated data provide the advantage of evaluating scMappR while controlling for differential expression between conditions, cell-type specificity between simulated cell-types and the bulk RNA-seq sample's cellular composition (Figure 2A–C). Briefly, we simulated cell-type specific RNA-seq data using five cell-types, two conditions and six replicates (60 samples total). Simulated genes could be cell-type specific DEGs, genes that are differentially expressed across all cell-types, cell-type markers and genes with no differential expression (Figure 2A). We varied the degree of cell-type specificity in different iterations of the simulation. We then simulated bulk RNA-seq data with two conditions and six replicates (Figure 2B). We varied the cell-type proportions of the simulated bulk samples in different iterations of the simulation. We measured cell-type specific differential expression, bulk differential expression, and cwFold-changes of the bulk DEGs. With each parameter in cell-type specificity, cell-type proportion, and batching strategy (Supplementary Tables S1 and S2), we evaluated how accurately cwFold-changes sort into cell-types with cell-type specific differential expression and if the distribution of cwFold-changes are more highly correlated to cell-type specific DEGs than the distribution of bulk DEGs alone (Figure 2C and Supplementary Table S1).

We first tested the ability of scMappR to correctly map DEGs when altering the degree of cell-type specificity between cell-types. We found that scMappR was able to assign DEGs to the correct cell-type with a false positive rate $<5\%$ (0–2.2%, Supplementary File S2) regardless of cell-type proportion and cell-type specificity (Figure 2D; Supplementary Figure S2A,B and File S2). We were only able to assign DEGs to cell-types if we set the fold-change of a cell-type marker to be >2, suggesting that our true positives are being driven by cell-type specific expression (Figure 2D). Similarly, the cwFold-changes were more highly correlated

to cell-type specific DEGs than bulk DEGs alone (Figure 2E and Supplementary Figure S2C). These findings were consistent within the bulk sample in cell-types that were uncommon (5–7% of bulk sample) and moderately common (20–23% of bulk sample; Figure 2D,E and Supplementary Figure S2A,B).

Our simulations interrogated if scMappR can detect DEGs when we fix cell-type specificity ($s = 5$) and vary cell-type proportions within a bulk sample (Supplementary Tables S3 and S4). We found that scMappR maintained similar ratios between true and false positives at all cell-type proportions except for the rarest cell-type (1/15 of the bulk sample) that had fewer true positives (Figure 2F). The decrease in the true positive rate of the rarest cell-type is likely driven by fewer cell-type specific DEGs being detected as being differentially expressed in the bulk sample. Similarly, the cwFold-changes of simulated bulk DEGs were more highly correlated to simulated cell-type specific DEGs compared to the simulated bulk DEGs alone regardless of cell-type proportion (P paired t-test $= 3.91 \times 10^{-4}$, Rho increase $= 0.0848$). (Figure 2G). To determine the importance of estimated cell-type proportion on correctly assigning DEGs to their cell-type, we tested a null case where the entire bulk sample originated from cell-type 1. In this instance, the other four cell-types not contained in the bulk sample had estimated cell-type proportions between 4.04 and 6.34%, and a false positive rate of 7.12% (Supplementary Figure S2D). This analysis suggests that scMappR can be used to test the purity of cell-type specific RNA-seq.

To investigate the implications that batch effects in the bulk RNA-seq data could have on scMappR results, we simulated bulk RNA-seq data with a batch effect that is weaker than the condition effect, and a batch effect that is stronger than the condition effect. Overall, we found that in both the weaker and stronger batch effect, we did not find a significant increase in false positives or false negatives compared to when there was no simulated batch effect in the bulk RNA-seq data (Supplementary Figure S3A,C). When using a weaker batch effect, we found that the distribution of cwFold-changes remained more highly correlated to cell-type specific DEGs than bulk DEGs alone (P-paired t-test $= 0.00175$, Rho increase $= 0.0549$) (Supplementary Figure S3B). When using a stronger batch effect, the distribution of cwFold-changes was not more highly correlated to cell-type specific DEGs than bulk DEGs alone (P-paired t-test $= 0.111$, Rho increase $= 0.0129$; Supplementary Figure S3D).

DEG lists re-ranked by scMappR reflect cell-type-specific differential expression

An ideal benchmark for scMappR is a dataset that contains RNA-seq of a bulk tissue, with at least two conditions, where DEGs can be calculated from the heterogeneous bulk tissue, as well as from the purified cell-types comprising the bulk tissue. Monaco *et al.*, 2019 produced such a dataset consisting of bulk RNA-seq in peripheral blood mononuclear cells (PBMC) ($N = 13$) together with RNA-seq of 29 cell-types ($N = 4$ cell-type) isolated from PBMCs (27). This dataset avoids some of the inherent limitations of calculating DEGs with scRNA-seq alone such as gene dropout and batch effects due to the limited number of scRNA-seq replicates per run (6,7). Monaco *et al.*, 2019 used males and females in the bulk RNA-seq ($N = 9$ male, 4 females) and in the cell-sorted RNA-seq analyses ($N = 2$ males, 2 females), allowing for cell-type specific measurements of sex differences (27) (Figure 3A, B). To begin, we built a signature matrix using these cell-sorted RNA-seq data by calculating differential expression of each cell-type (sexes combined) versus all others cell-types with DESeq2 (2) (see 'Materials and Methods' section).

Through bulk RNA-seq analysis, we identified 59 DEGs between sexes in PBMCs (Wald's test; FDR adjusted P-value < 0.05 and absolute fold-change > 1.5) (Supplementary Figure S3A). RNA-seq deconvolution tools including DeconRNA-seq function optimally when there are more samples than cell-types (16,61). Therefore, we tested scMappR with the top 12 most variable cell-types (one fewer than the 13 bulk samples). We then tested if the fold-changes of these 59 sex-biased DEGs were more highly correlated to the same 59 DEGs genes in these 12 cell-types using Spearman's correlation (Figure 3C). We found that for every cell-type, scMappR's cwFold-changes had a higher or equal correlation with cell-type specific DEGs than the bulk correlation with cell-type specific DEGs (average rho increase $= 0.0471$, one-tailed paired Student's t-test, $P = 2.00 \times 10^{-4}$) (Supplementary File S3). Applying the 'cwFoldChange_evaluate' function to the 59 sex-biased DEGs identified 40 genes with cell-type specificity in at least one cell-type (Supplementary Figure S4). On average, 80% of the DEGs assigned to each cell-type have an absolute fold-change > 1.5 in that cell-type (mean assignment $= 80.0\%$, sd assignment $= 18.9\%$) (Supplementary File S4). Overall, scMappR significantly increased the correlation between bulk DEGs and cell-type specific DEGs in a study that already contained a high correlation between cell-type specific DEGs and bulk cell-type specific differential expression (Rho $= 0.535$–0.777).

We next tested whether genes that are differentially expressed in every cell-type falsely influence the cell-type specificity of cwFold-changes. Genes mapping to the Y-chromosome are inherently male biased, and XIST and TSIX are inherently female biased. We removed these 18 DEGs, re-calculated cwFold-changes and re-calculated the correlation between cwFold-changes and cell-type specific fold-changes versus bulk fold-changes versus cell-type specific fold-changes. We found the overall correlation between cwFold-changes and cell-type specific DEGs was still higher than the correlation between bulk DEGs and cell-type specific DEGs (average rho increase $= 0.0670$, one-tailed paired Student's t-test, $P = 0.00844$), showing that ubiquitously expressed DEGs do not improperly influence scMappR's cwFold-changes.

We next tested whether scMappR is robust to any combination of cell-types, and not just the 12 most variable cell-types. Monaco *et al.*, 2019 contained 13 bulk RNA-seq (PBMC) samples, allowing us to test scMappR using 12 cell-types at once (16,61). We randomly sampled 12/29 cell-types and re-calculated the P-value and change in correlation for 100 permutations to ensure that our results are not biased by the cell-types that we selected. This permutation-based analysis showed that regardless of the

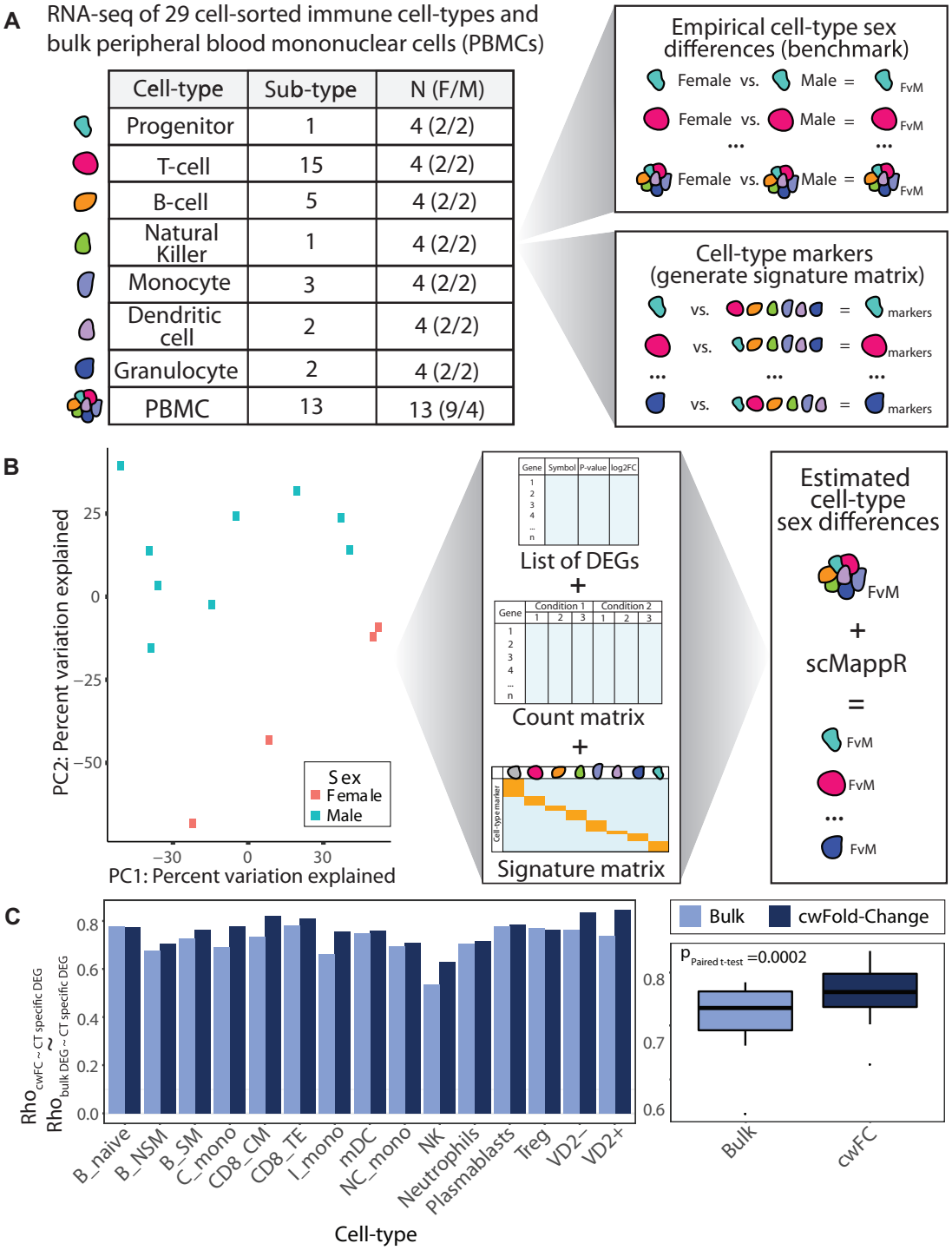

Figure 3. Benchmarking scMappR workflow and results. (**A**) Overview of samples and cell-types from Monaco *et al.*, 2019. Sex differences within each cell-type are computed and the cell-type specific fold-changes in the genes that are differentially expressed in the peripheral blood mononuclear cells (PBMC) dataset are used. Each column of the signature matrix is the fold-change of expression from each cell-type against all the other cell-types and each row is a cell-type marker. (**B**) Overview of how scMappR was used to estimate cell-type specific sex differences from PBMCs. Principal component analysis shows linear separation of male and female PBMC samples. Differentially expressed genes derived from computing sex differences, the normalized count matrix, and signature matrix generated in (A) were imported into scMappR. (**C**) Improvement that cell-weighted fold-changes (cwFold-changes) have on cell-type specificity for every cell-type measured with a bar chart. Dark bars are the correlation cwFold-changes with cell-type specific fold-changes. Light = bars are the correlation between cell-types (left) and a boxplot of the correlations across cell-types (right). Improvement in correlation is measured with a one-tailed paired Student's *t*-test; Bulk/PBMC, Peripheral Blood Mononuclear Cells; Neutrophils, Neutrophils; Progenitor, Progenitor; Basophils, Basophils; pDC, Plasmacytoid dendritic cells; Plasmablast, Plasmablast; mDC, myeloid dendritic cells; B_naïve, naïve B cells; NC_mono, nonclassical monocytes; C_mono, classical monocytes; MAIT, MAIT cells; B_SM, Switched memory B cells; VD2-, non-Vd2 gd T-cells.

cell-types selected, there was always a statistically significant increase in cell-type specificity (mean P-value = 1.83×10^{-4}, mean Rho increase = 0.0545) (Supplementary Figure S5). scMappR's cwFold-changes had a higher correlation to cell-type specific DEGs than bulk DEGs had with cell-type specific DEGs for two reasons. First, scMappR increased the rank of differentially expressed cell-type markers (Supplementary Figure S5). Second, scMappR decreased the rank of DEGs that were not expressed in a particular cell-type (Supplementary Figure S5). Together, this analysis showed that scMappR can significantly improve the correlation of bulk DEGs to cell-type specific DEGs.

scMappR reveals cell-type specific DEGs during mouse kidney regeneration

After benchmarking scMappR, we tested how scMappR can be used to identify cell-types that contribute to DEGs generated from a representative, well-designed bulk RNA-seq study of a heterogeneous tissue. To do this, we reanalyzed data from Valle Duraes *et al.*, 2020 (14) who investigated gene expression changes involved in mouse kidney regeneration before and after injury (14). Kidney regeneration involves multiple cell-type specific processes (62–65), and importantly Valle Duraes *et al.*, 2020 used bulk RNA-seq in conjunction with histopathology, cell sorting, and scRNA-seq to implicate T-Cell recruitment as a critical part of the regeneration process (14). Valle Duraes *et al.*, 2020 used a bulk RNA-seq study design (14) that contains 50 total samples split into fibrosis (using wild-type mice) and regeneration (using B6.Cg-Foxp3tm2(EGFP)Tch/J mice) models after injury (days 0, 3, 7, 14, 28 and 42) (N = 3–4 per condition/timepoint) (Supplementary Figure S6). We reasoned that this is an ideal model RNA-seq study to test scMappR as Valle Duraes *et al.*, 2020 is well-powered, and includes detailed experimental follow-up of cell-type specific responses (14).

For simplicity, we focused on the comparison of the initial two timepoints as these contained the most dramatic changes (day 0 ('naïve') versus day 3 (injury induced 'regeneration'). For every comparison, all samples were used in the RNA-seq deconvolution step of scMappR's generation of cwFold-changes (all time periods in regeneration and fibrosis). In conjunction, a kidney scRNA-seq dataset from *Tabula Muris*, 2018 (28) was preprocessed and stored in scMappR. We then used scMappR to identify which cell-types are involved in kidney regeneration using both bulk and scRNA-seq datasets.

After reprocessing data in Valle Duraes *et al.*, 2020 (14), we identified 2855 DEGs between the 'naïve' and 'regeneration' groups. We found that 394 of these DEGs were kidney cell-type markers in *Tabula Muris,* 2018 (28) (Figure 4A). Using scMappR, we then asked which cell-types had the highest cwFold-changes in DEG comparisons between naïve day 0 and regeneration day 3 groups in the whole kidney. We found clear signatures of fibroblasts, smooth muscle, and endothelial cells, all of which have well-documented roles in kidney regeneration (62–65) (Figure 4B). A subset of immune ('Macrophage, dendritic') specific DEGs were also found (Figure 4B, Table 1). The immune-specific DEGs were less prevalent than other cell-types (Figure 4B), likely due to a lower proportion of immune cells in the kidney (66).

The immune 'Macrophage, dendritic' cell-type contains 430 cell-type markers that enrich for many immune related processes (immune system processes: precision = 0.537, one-tailed hypergeometric test FDR adjusted P-value = 3.65×10^{-87}; innate immune response: precision = 0.223, one-tailed hypergeometric test FDR adjusted P-value = 1.78×10^{-38}; adaptive immune response: precision = 0.184, one-tailed hypergeometric test FDR adjusted P-value = 7.95×10^{-36}; T-cell activation: precision = 0.161, one-tailed hypergeometric test FDR adjusted P-value = 2.10×10^{-32}). Furthermore, the original *Tabula Muris, 2018* study labeled this cell-type population as 'Macrophage and Natural Killer' (28). Interestingly, many cells within this population contain a high expression of naïve T-cell markers like *Ccr7* and *Nkg7* (14,28). These results are unsurprising as T-cells are present in the uninjured kidney (67). Therefore, although this cluster was given the 'Macrophage, dendritic' label, it might be better interpreted as a cell-type representing the heterogeneous immune-cell population in the *Tabula Muris,* 2018 (28) kidney.

Overall, the top five most significant pathways of these re-ranked DEGs showed a common regeneration phenotype across different cell-types at the pathway level (Supplementary Figure S7). For each cell-type, between 52 and 59% of the pathways were shared between the enriched pathways derived from bulk differential expression compared to pathways derived from genes re-ranked by cwFold-changes (Supplementary Table S6). Pathways that were only identified in the cell-type specific pathway analyses but not in bulk pathway enrichment were biologically relevant. One such pathway is the 'Immune System' gene ontology, which was not significantly enriched with the bulk DEG list but was highly enriched when re-ranking the same DEGs but by their 'Macrophage, dendritic' cwFold-changes (FDR adjusted P-value = 1.62×10^{-6}). The top five most significant pathways identified by ordering genes based on their rank-change between bulk DEGs and cwFold-changes (Supplementary Figure S8) were related to their cell-type, including significant enrichment of immune related pathways in the 'Macrophage, dendritic' cell-type (immune response: precision = 0.125, intersection of DEGs and pathway = 156 genes, one-tailed hypergeometric test FDR adjusted P-value = 6.67×10^{-19}, cytokine–cytokine receptor interaction: precision = 0.0320, intersection of DEGs and pathway = 37 genes, one-tailed hypergeometric test FDR adjusted P-value = 3.01×10^{-7}, phagosome: precision = 0.0240, intersection of DEGs and pathway = 30 genes, one-tailed hypergeometric test FDR adjusted P-value = 1.82×10^{-5}, and phagocytosis: precision = 0.126 intersection of DEGs and pathway = 30 genes, one-tailed hypergeometric test FDR adjusted P-value = 3.32×10^{-5}) (Figure 4C). Taken together, scMappR increases the rank of cell-type specific DEGs, thus allowing for biologically relevant cell-type specific pathway analysis.

We then investigated the potential biological pathways that the 2855 DEGs between naïve (day 0) and regeneration (day 3) were assigned to the immune specific 'Macrophage, dendritic' cell-type. Of the 2855 bulk DEGs between naïve (day 0) and regeneration (day 3), 431 were cell-type specific

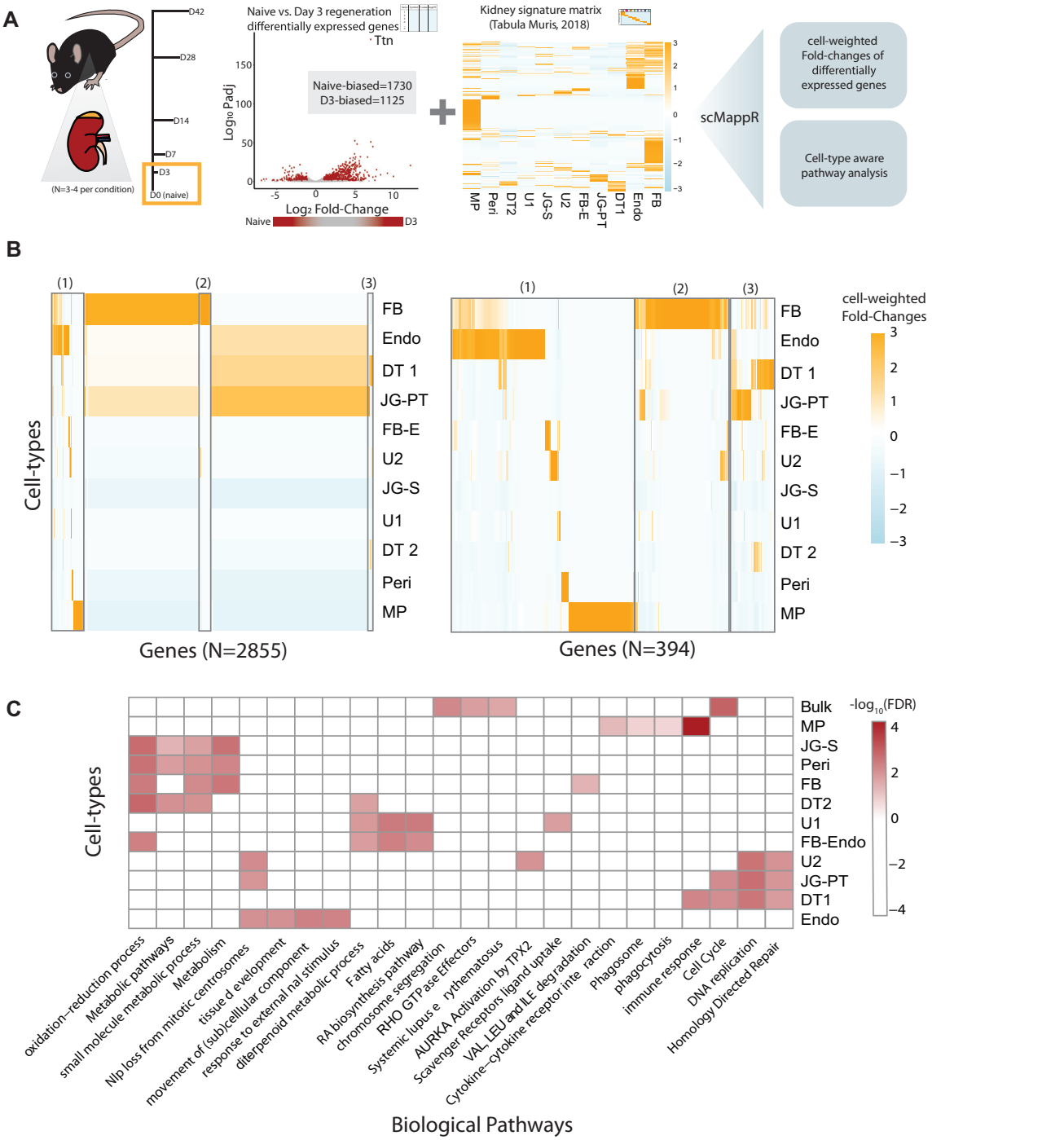

Figure 4. Application of scMappR to identify which cell-types are responsible for differentially expressed genes in kidney regeneration. (**A**) Valle Duraes *et al.*, 2020 completed RNA-seq of C57BL/6J mice kidneys at naïve (day 0) and multiple timepoints of kidney regeneration post-injury. Between naïve and regeneration day 3 comparisons (shown here), we identified 2855 significantly differentially expressed genes. We then used scMappR to compute cwFold-changes. The normalized count data, the list of differentially expressed genes, and a signature matrix were inputs for this analysis. We used normalized count data from all samples, differentially expressed genes from naïve versus kidney regeneration (naïve (day 0) versus day 3 comparison shown here), and the signature matrix from scRNA-seq in the kidney completed by *Tabula Muris, 2018.* (**B**) Heatmap of gene normalized cwFold-changes of all 2855 differentially expressed genes (left) and the 394 differentially expressed genes that are also identified as cell-type markers in *Tabula Muris*, 2018 (right). The heatmaps on the left and right were produced in the same way except that in the heatmap on the right the genes are filtered for cell-type markers in *Tabula Muris*, 2018. (**C**) A cell-type normalized matrix of the top four most enriched pathways from cell-type specific pathway analysis. For each cell-type, genes were re-ranked by their increase in cell-type specificity before pathway analysis was completed. Bulk, bulk kidney; MP, Macrophage, Dendritic; JG-S, Juxtaglomerular, Stem; Peri, Pericyte; FB, Fibroblast; DT2, Distal Tubule 2; U1, Unknown 1; FB-Endo, Fibroblast-Endothelial; DT1, Distal Tubule 1; JG-PT, Juxtaglomerular, Proximal tubule; U2, Unknown 2; Endo, Endothelial.

Table 1. Over-representation of cell-type markers of consistently processed scRNA-seq data in over 100 mouse tissues when testing 34 T-cell markers

SRA ID	Tissue	Label from CellMarker database	Label from Panglao database	Number of cell-type markers	Number of overlapping cell-type markers	Odds ratio	Adjusted *P*-value
SRA653146	Trachea	Lymphocyte	Nuocytes	122	15	83.1	2.47×10^{-18}
SRA667466	Cortex 3	Lymphocyte	Nuocytes	122	15	83.1	2.47×10^{-18}
SRA653146	Muscle	Myeloid cell	Natural killer cells	88	13	90.6	1.08×10^{-16}
SRA667466	Dorsal midbrain	Myeloid cell	Natural killer cells	88	13	90.6	1.08×10^{-16}
SRA748166	Cardiac tissue	T cell	Natural killer cells	132	13	60.2	9.58×10^{-15}
SRA801845	Cardiac progenitor cells	T cell	Natural killer cells	132	13	60.2	9.58×10^{-15}
SRA638923	Small intestine	Immune cell	Natural killer cells	92	11	66.9	4.40×10^{-13}
SRA711739	Embryonic fibroblasts	Epithelial cell	Natural killer T cells	60	10	89.5	4.40×10^{-13}
SRA757237	Bone marrow	Epithelial cell	Natural killer T cells	60	10	89.5	4.40×10^{-13}
SRA653146	Spleen	T cell	Thymocytes	68	10	78.9	1.03×10^{-12}

in at least one cell-type and 165 DEGs were assigned to the immune specific 'Macrophage, dendritic' cell-type. The most enriched pathway for these DEGs is the 'immune response' pathway (immune response: precision = 0.475, intersection of DEGs and pathway = 66 genes, one-tailed hypergeometric test FDR adjusted *P*-value = 5.73×10^{-36}) and 28 genes were associated with T-cell activation (T-cell activation: precision = 0.178, intersection of DEGs and pathway = 28 genes, one-tailed hypergeometric test FDR adjusted *P*-value = 3.24×10^{-15}). Some of these genes include immune-system regulators involved in kidney fibrosis and regeneration including *Ccr7*, *Ccr2*, *Vista* and *Tgfb1* (62,68–70).

In their original manuscript, Valle Duraes *et al.,* 2020 FACS-sorted (71) T-cells and T-regulatory cells (14) with the naïve and regenerating kidney samples before completing RNA-seq and differential analysis (Figure 5A). These experiments allowed us to investigate the extent that their bulk DEGs that scMappR assigned to the immune cell-type overlapped with DEGs directly measured in T-cells and T-regulatory cells (Figure 5B) (Supplementary File 5). Of the 2855 bulk DEGs between naïve (day 0) and regeneration (day 3), 431 were cell-type specific in at least one cell-type and 165 DEGs were assigned to the immune specific 'Macrophage, dendritic' cell-type. Of these 165 DEGs, 28 were associated with T-cell activation (T-cell activation: precision = 0.178, intersection of DEGs and pathway = 28 genes, one-tailed hypergeometric test FDR adjusted *P*-value = 3.24×10^{-15}). Some of these genes include immune-system regulators involved in kidney fibrosis and regeneration including *Ccr7*, *Ccr2*, *Vista* and *Tgfb1* (62,68–70). We then compared the 165 immune-specific 'Macrophage, dendritic' mapping DEGs to the DEGs measured in T-cells and T-regulatory cells directly. We found that 81 DEGs overlapped between the bulk DEGs mapped to immune cells (165 genes) and DEGs measured in T-cells (3074 genes) directly (odds ratio = 2.76, FDR adjusted *P*-value = 3.87×10^{-11}) (Figure 5C). This overlap is primarily driven by regeneration biased DEGs Figure 5D). We found that 29 DEGs overlapped between the bulk DEGs mapped to immune cells and DEGs measured in T-regulatory cells directly (Odds Ratio = 2.35 FDR adjusted *P*-value = 5.23×10^{-4}) (Figure 5D). Similar to what was observed for T-cells, this overlap is primarily driven by regeneration-biased DEGs (Figure 5F). Together, we showed that scMappR assigned specific bulk DEGs to the correct cell-type in real RNA-seq data.

We evaluate how bulk RNA-seq sample size, RNA-seq deconvolution and scRNA-seq normalization can influence cwFold-changes (see Supplementary Methods and Supplementary Figure S9). Briefly, by re-calculating cwFold-changes with seven total samples, we found that scMappR calculated cwFold-changes with the same rank-order of DEGs within and across cell-types if there are more cell-types than samples (Supplementary Table S7). We found that in this dataset RNA-seq deconvolution using DeconRNA-seq, WGCNA and DCQ had no influence on the order of DEGs within a cell-type. These differences in RNA-seq deconvolution approaches did influence the rank-order of cell-types within a DEG because of the influence that RNA-seq normalization and RNA-seq deconvolution have on cell-type proportions. scRNA-seq processing with Seurat V3 versus scTransform inherently influenced how scMappR would calculate cwFold-changes because they identified a different number of clusters (i.e. cell-types). In conclusion, scMappR allows for DEGs normalized RNA-seq data, and processed scRNA-seq data from any method, how these data are pre-processed can influence the generated cwFold-changes. Proper data-preprocessing is therefore important for the most accurate results.

scMappR: projection of a generic gene list onto scRNA-seq data

We also designed scMappR as tool that can complete traditional enrichment gene-set enrichment of a list of genes (e.g. a list of putative genes uncovered by a genome wide association study for a complex trait). We curated all of the cell-types and cell-type markers from the 331 signature matrices into a gene-set database. The 'tissue_by_celltype_enrichment' function allows for gene set enrichment of a gene list on scMappR's curated gene-set database. This gene-set database has the advantage of using every cell-type marker originating from a consistent bioinformatic analysis. Alternatively, 'tissue_scMappR_internal' and 'tissue_scMappR_custom' provide a more hypothesis driven approach where users can ask if their list of genes are more likely to be expressed in one cell-type compared to other cell-types in the same tissue based on the over-representation of cell-type markers.

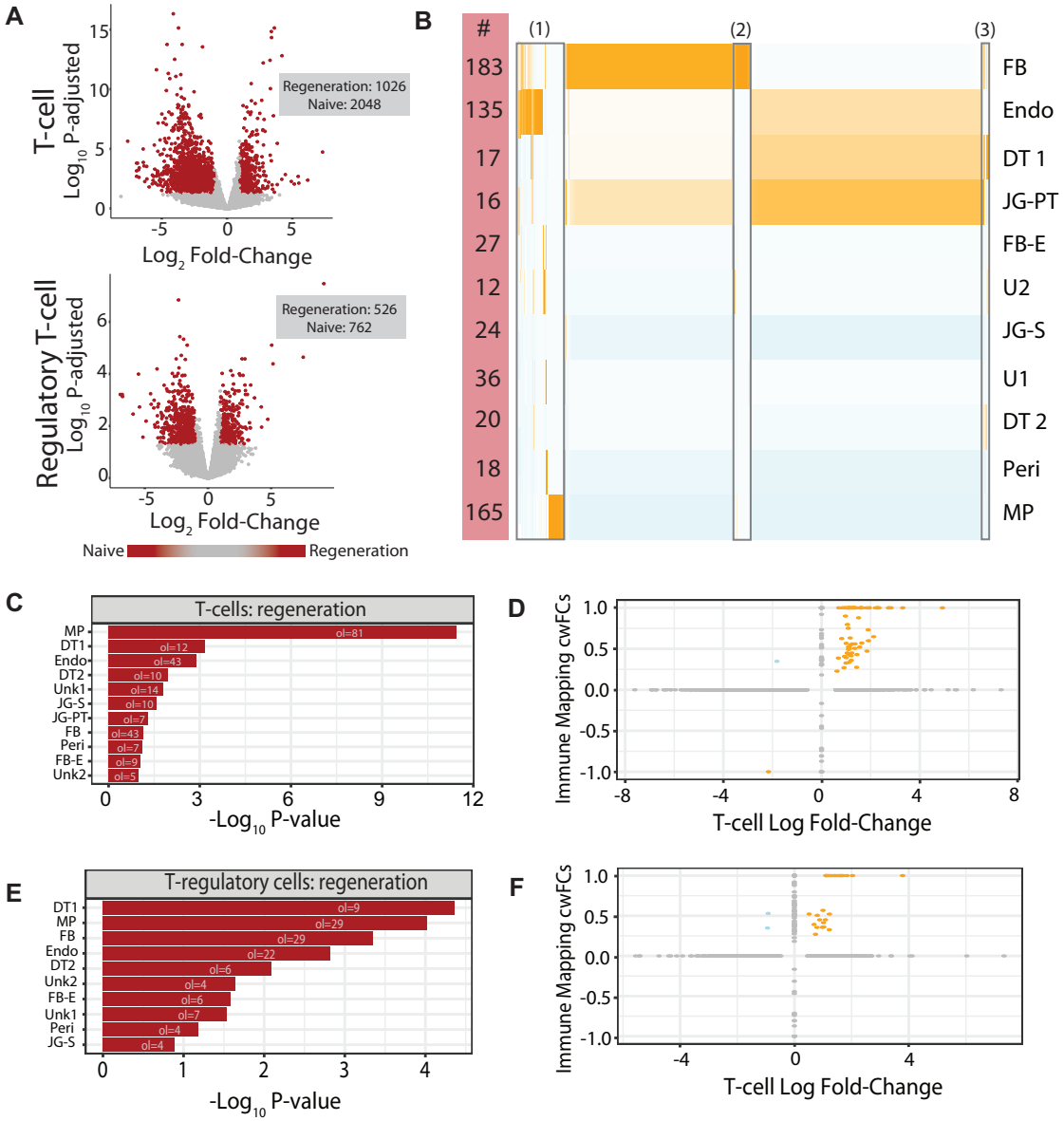

Figure 5. Comparison of bulk DEGs involved in kidney regeneration mapping to immune cells by scMappR with DEGs involved in kidney regeneration identified from FACS sorted T-cells and T-regulatory cells. (**A**) Volcano plots of DEGs between a naïve and regenerating kidney in FACS sorted T-cells and T-regulatory cells. (**B**) Overview of cwFold-changes in 2855 DEGs. Rows are cell-types and columns are DEGs. Numbers on the left side of the heatmap are the number of DEGs significantly mapping to each cell-type. (**C**) Enrichment of DEGs identified in T-cells on DEGs mapping to each cell-type. (**D**) Scatterplot of the cwFold-changes mapping to the 'Macrophage, Dendritic' cell-type (*y*-axis) and DEGs from FACS-sorted T-cells. (**E**) Enrichment of DEGs identified in T-regulatory cells on DEGs mapping to each cell-type. (**F**) Scatterplot of the cwFold-changes mapping to the 'Macrophage, Dendritic' cell-type (*y*-axis) and DEGs from FACS-sorted T-regulatory cells. Bulk, bulk kidney; MP, Macrophage, Dendritic; JG-S, Juxtaglomerular, Stem; Peri, Pericyte; FB, Fibroblast; DT2, Distal Tubule 2; U1, Unknown 1; FB-Endo, Fibroblast-Endothelial; DT1, Distal Tubule 1; JG-PT, Juxtaglomerular, Proximal tubule; U2, Unknown 2; Endo, Endothelial.

In addition to disentangling the cell-type specific role of bulk DEGs, scMappR can facilitate the understanding of cell-type specific expression in any list of genes. We tested the cell-type enrichment for the 2855 DEGs measured between naïve day 0 and regeneration day 3 in the kidney across all the cell-types and cell-type markers stored in scMappR (see 'Materials and Methods' section). The top ten most significantly enriched cell-types were 'proliferating cells and gamma delta T cells' (Supplementary Table S8). To characterize the $CD4^+$ scRNA-seq dataset, Valle Duraes *et al.*, 2020 (14) utilized a curated set of 34 T-cell marker genes (72). We asked if scMappR in combination with our uniformly processed scRNA-seq data would also consider these as T-cell marker genes. Of the top ten most enriched cell-types, all ten were immune cell-types and four out of ten were from cell-types labelled as T-cells (Table 1).

In addition to testing lists of genes across compendiums of scRNA-seq data, scMappR is useful for interrogating a specific, biologically relevant tissue. This approach is valuable when users have a list of genes from a particular

Table 2. Over- and under-representation of kidney cell-type markers from scRNA-seq data generated by *Tabula Muris,* 2018 when testing 34 T-cell markers

Cell Type	Total number of cell-type markers in *Tabula Muris, 2018*	Odds ratio	Adjusted *P*-value	Number of genes	T-cell marker genes
Macrophage, dendritic	430	20.9	0.00115	10	Klf2, Rgs2, Ccl4, Cd83, Nkg7, C cl5, Ccr7, Sell, Ifng, Cd7
Endothelia	560	0.138	0.169	1	Klf2
Fibroblasts	548	0.317	0.793	2	Klf2, Gata3
Distal tubule	48	2.67	0.927	1	Gata3
Proximal tubule, juxtaglomerular	23	0.00	1.00	0	
Distal tubule1	111	0.00	1.00	0	
Unknown	49	0.00	1.00	0	
Pericyte	49	0.00	1.00	0	
Fibroblast, endothelia	43	0.00	1.00	0	
Unknown1	3	0.00	1.00	0	
Stem, juxtaglomerular	23	0.00	1.00	0	

tissue but cell-type proportions cannot be integrated with scRNA-seq expression (e.g. genes mapping to ChIP-seq peaks) (73,74). As an example, we compared the 2855 DEGs between naïve kidney and kidney regeneration (3-h post injury) against the *Tabula Muris*, 2018 (28) kidney scRNA-seq study. We found an over-representation of the immune ('Macrophage Dendritic') cell-type in the upregulated (regeneration biased) DEGs (FDR adjusted *P*-value = 1.43×10^{-5}, odds-ratio = 1.86) and an under-representation of the same cell-type in the downregulated (naïve biased) DEGs (FDR adjusted *P*-value = 4.23×10^{-5}, odds-ratio = 0.33) (Supplementary Table S9). Since the 34 T-cell markers exclusively enriched for the immune ('Macrophage, dendritic') cell-type (FDR adjusted *P*-value = 0.00115, odds-ratio = 20.9) (Table 2), we suggest that scMappR did detect evidence of T-cell infiltration which Valle Duraes *et al.*, 2020 experimentally validated in their study (14). Overall, our results show that scMappR can calibrate genes from a representative RNA-seq study design and detect biologically relevant cell-type specific enrichments from gene lists using compendiums of scRNA-seq data.

DISCUSSION

scMappR is an R package designed for the primary purpose of estimating which cell-types contribute to a list of DEGs from bulk RNA-seq. scMappR integrates both cell-type expression and cell-type proportion to generate cell-type specificity scores (cwFold-changes). We showed using simulated and real RNA-seq data that scMappR correctly assigns bulk differentially expressed to a cell-type(s) where the gene is differentially expressed (Figures 2, 4 and 5). Re-ranking differentially expressed genes by their cwFold-changes reflects the distribution of DEGs within a cell-type and can also improve cell-type specific pathway analysis. We showed that the distribution of cwFold-changes are more similar to cell-type specific DEGs in both simulated and real data (Figures 2 and 3). Computing cwFold-changes on DEGs across kidney regeneration allowed for the evaluation of the cell-types from the measured bulk DEGs and for cell-type specific pathway analysis (Figures 4 and 5). scMappR can be performed in many experimental contexts and should provide valuable cell-type specificity to a list of DEGs.

The general usability of scMappR with bulk RNA-seq analysis is facilitated in two ways. First, scMappR stores consistently processed mouse and human signature matrices for users to choose from. Second, scMappR contains the bioinformatic pipelines that allow users to reprocess any scRNA-seq count dataset into a signature matrix. There are thousands of viable scRNA-seq processing pipelines (75), and to accommodate this, scMappR allows users to input their own signature matrix, scRNA-seq count data or processed scRNA-seq dataset. From there, scMappR has functions to convert this data into a signature matrix compatible with scMappR's cwFold-change generation.

ScMappR improves the cell-type specificity of DEGs measured with traditional bulk RNA-seq analysis, but it is not a tool designed to detect *de novo* DEGs. Methods such as BSeq-sc (25) and csSAM (4) aim to identify *de novo* DEGs that were not originally detected in bulk RNA-seq by leveraging estimated cell-type proportions, but not cell-type specificity. BSeq-sc (25) identifies *de novo* DEGs by using estimated cell-type proportions as a covariate of differential analysis before applying csSAM, a least-squares regression and empirical false discovery rate (4). BSeq-sc (25) thus requires a larger sample size (e.g. 82 samples for three cell-types) to detect *de novo* DEGs. In contrast, scMappR is designed to work with typically RNA-seq study designs with the main caveat being that the number of samples used needs to exceed the number of cell-types used for the deconvolution analysis. If users need to pre-select cell-types for scMappR, picking cell-types that are abundant within the bulk sample and cell-types with large differences in gene expression from one another may lead to a higher accuracy of mapped cwFold-Changes (Figure 2).

One limitation of scMappR is that it must use signature-matrix based RNA-seq deconvolution tools to complete the normalization step in generating cwFold-changes. Signature matrix-based RNA-seq deconvolution tools are also convenient for generating cwFold-changes because of their computational efficiency. When calculating cwFold-changes, the number of times we estimate cell-type proportions is the number of DEGs + 1. We provide two ways to evaluate if the RNA-seq deconvolution completed in scMappR is compatible with the user's bulk RNA-seq experiment. First, the 'compare_deconvolution_methods' function evaluates the dif-

ferences in cell-type proportions between the three RNA-seq deconvolution tools in scMappR DeconRNA-seq (16), DCQ (29) and WGCNA (30), and optionally user supplied cell-type proportions. Second, cell-type proportions from any nonsignature matrix based deconvolution method (e.g. xCell, cell-population mapping) (34,35) or empirically measured cell-type proportions may be used in the 'cwFoldChange_evaluate' function (Equation 12). This way, a cwFold-change can only be assigned to a cell-type if its cell-type specificity passes all thresholds and is greater than the researcher supplied cell-type proportions.

scMappR leverages scRNA-seq data to characterize the cell-type specificity of a list of bulk DEGs while providing a cell-type marker database to test the over-representation of cell-type markers in any gene list. Currently, single-cell genomic technologies are evolving and expanding to include new assays such as a single-cell assay for transposable accessible chromatin sequencing (scATAC-seq). (76,77) and single cell DNA methylation (DNAm) (78,79), scRNA-seq across many biological conditions with replicates, and single-cell genomics with fewer technical limitations. As these methodologies improve, tools like scMappR that aid in integrating bulk and single-cell differential genomics will become increasingly important.

In summary, we have shown that scMappR can accurately estimate which cell-types contain DEGs identified from bulk data. scMappR also has the potential to uncover biological signals that may have otherwise been masked in traditional bulk differential analysis. The scMappR method is stored in a user-friendly R package that provides supplementary pipelines to support users with diverse experimental designs and sample sizes. Overall, scMappR should be easy to incorporate into existing RNA-seq pipelines and serve as a facile way to incorporate scRNA-seq data into differential gene expression analyses.

DATA AVAILABILITY

The scMappR R package is available at CRAN (stable release) https://cran.r-project.org/web/packages/scMappR/index.html. The scMappR developmental version is available on GitHub https://github.com/wilsonlabgroup/scMappR_Data. All code and files to generate figures and tables can be found on figShare (https://figshare.com/s/3b5cfb597a0b3bc2801c).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

Author Contributions: D.S., M.F.M., A.G. and M.W. conceived of the method described in this manuscript. D.S. created the R code and the R package with support from H.Z. Troubleshooting and methodological utility and testing of scMappR tool was completed by D.S., M.F.M., L.E., H.H. and C.C. The manuscript was written by D.S., M.F.M., C.C., A.G. and M.W. with support from all authors. M.H., A.G. and M.W supervised the work. We would also like to thank Dr Aziz Mezlini for his advice in simulating RNA-seq data, Erik Drysdale in how to formulate the methodology, and Dr Marla Sokolowski for her advice in making the manuscript interpretable by a nonspecialized audience.

FUNDING

Canadian Network for Research and Innovation in Machining Technology, Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN-2019–07014 to M.W.]; Ontario Ministry of Research, Innovation and Science (to M.W., A.G.); NSERC CGS M, PGS D and Ontario Graduate Scholarships (to D.S.); NSERC PGS D, the association computing machinery special interest group on high performance computing (ACM/SIGHPC) Intel Computational and Data Science Fellowship (to M.F.M.); CIHR [202003PJT-437197 to M.M.H., M.W.]; SickKids (to C.C.); Genome Canada Genomics Technology Platform, The Centre for Applied Genomics (to H.H.); NSERC [RGPIN 2018–04780, RGPAS 2018–522465 to M.M.H.]; Ontario Early Researcher Award.
Conflict of interest statement. None declared.

REFERENCES

1. Stark,R., Grzelak,M. and Hadfield,J. (2019) RNA sequencing: the teenage years. *Nat. Rev. Genet.*, **20**, 631–656.
2. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
3. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
4. Shen-Orr,S.S., Tibshirani,R., Khatri,P., Bodian,D.L., Staedtler,F., Perry,N.M., Hastie,T., Sarwal,M.M., Davis,M.M. and Butte,A.J. (2010) Cell type-specific gene expression differences in complex tissues. *Nat. Methods*, **7**, 287–289.
5. Soneson,C. and Robinson,M.D. (2018) Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods*, **15**, 255–261.
6. Wang,T., Li,B., Nelson,C.E. and Nabavi,S. (2019) Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics*, **20**, 40.
7. Lähnemann,D., Köster,J., Szczurek,E., McCarthy,D.J., Hicks,S.C., Robinson,M.D., Vallejos,C.A., Campbell,K.R., Beerenwinkel,N., Mahfouz,A. *et al.* (2020) Eleven grand challenges in single-cell data science. *Genome Biol.*, **21**, 31.
8. Huang,S., Sheng,X. and Susztak,K. (2019) The kidney transcriptome, from single cells to whole organs and back. *Curr. Opin. Nephrol. Hypertens.*, **28**, 219–226.
9. Mendizabal,I., Berto,S., Usui,N., Toriumi,K., Chatterjee,P., Douglas,C., Huh,I., Jeong,H., Layman,T., Tamminga,C.A. *et al.* (2019) Cell type-specific epigenetic links to schizophrenia risk in the brain. *Genome Biol.*, **20**, 135.
10. Tirosh,I., Izar,B., Prakadan,S.M., Wadsworth,M.H., Treacy,D., Trombetta,J.J., Rotem,A., Rodman,C., Lian,C., Murphy,G. *et al.* (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, **352**, 189–196.
11. Peng,T., Zhu,Q., Yin,P. and Tan,K. (2019) SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome Biol.*, **20**, 88.
12. Sosina,O.A., Tran,M.N., Maynard,K., Tao,R., Taub,M.A., Martinowich,K., Semick,S.A., Weinberger,D., Quach,B.C., Hyde,T.M. *et al.* (2020) Strategies for cellular deconvolution in human brain RNA sequencing data. bioRxiv doi: http://dx.doi.org/10.1101/2020.01.19.910976, 20 January 2020, preprint: not peer reviewed.
13. Wang,X., Park,J., Susztak,K., Zhang,N.R. and Li,M. (2019) Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.*, **10**, 380.

14. do,Valle Duraes, Lafont,A., Beibel,M., Martin,K., Darribat,K., Cuttat,R., Waldt,A., Naumann,U., Wieczorek,G., Gaulis,S. *et al.* (2020) Immune cell landscaping reveals a protective role for regulatory T cells during kidney injury and fibrosis. *JCI Insight*, **5**, e130651.
15. Chen,B., Khodadoust,M.S., Liu,C.L., Newman,A.M. and Alizadeh,A.A. (2018) Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol. Biol.*, **1711**, 243–259.
16. Gong,T. and Szustakowski,J.D. (2013) DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics*, **29**, 1083–1085.
17. Zhang,X., Lan,Y., Xu,J., Quan,F., Zhao,E., Deng,C., Luo,T., Xu,L., Liao,G., Yan,M. *et al.* (2019) CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.*, **47**, D721–D728.
18. Franzén,O., Gan,L.-M. and Björkegren,J.L.M. (2019) PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)*, **2019**, baz046.
19. Cao,Y., Zhu,J., Jia,P. and Zhao,Z. (2017) scRNASeqDB: A database for RNA-Seq based gene expression profiles in human single cells. *Genes (Basel)*, **8**, 368.
20. Abugessaisa,I., Noguchi,S., Böttcher,M., Hasegawa,A., Kouno,T., Kato,S., Tada,Y., Ura,H., Abe,K., Shin,J.W. *et al.* (2018) SCPortalen: human and mouse single-cell centric database. *Nucleic Acids Res.*, **46**, D781–D787.
21. Papatheodorou,I., Moreno,P., Manning,J., Fuentes,A.M.-P., George,N., Fexova,S., Fonseca,N.A., Füllgrabe,A., Green,M., Huang,N. *et al.* (2020) Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.*, **48**, D77–D83.
22. Rozenblatt-Rosen,O., Stubbington,M.J.T., Regev,A. and Teichmann,S.A. (2017) The Human Cell Atlas: from vision to reality. *Nature*, **550**, 451–453.
23. Schelker,M., Feau,S., Du,J., Ranu,N., Klipp,E., MacBeath,G., Schoeberl,B. and Raue,A. (2017) Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.*, **8**, 2032.
24. Khrameeva,E., Kurochkin,I., Han,D., Guijarro,P., Kanton,S., Santel,M., Qian,Z., Rong,S., Mazin,P., Sabirov,M. *et al.* (2020) Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains. *Genome Res.*, **30**, 776–789.
25. Baron,M., Veres,A., Wolock,S.L., Faust,A.L., Gaujoux,R., Vetere,A., Ryu,J.H., Wagner,B.K., Shen-Orr,S.S., Klein,A.M. *et al.* (2016) A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.*, **3**, 346–360.
26. Law,C.W., Chen,Y., Shi,W. and Smyth,G.K. (2014) voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.
27. Monaco,G., Lee,B., Xu,W., Mustafah,S., Hwang,Y.Y., Carré,C., Burdin,N., Visan,L., Ceccarelli,M., Poidinger,M. *et al.* (2019) RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Rep.*, **26**, 1627–1640.
28. Tabula Muris Consortium (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, **562**, 367–372.
29. Altboum,Z., Steuerman,Y., David,E., Barnett-Itzhaki,Z., Valadarsky,L., Keren-Shaul,H., Meningher,T., Mendelson,E., Mandelboim,M., Gat-Viks,I. *et al.* (2014) Digital cell quantification identifies global immune cell dynamics during influenza infection. *Mol. Syst. Biol.*, **10**, 720.
30. Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.
31. Zhu,L., Lei,J., Devlin,B. and Roeder,K. (2018) A unified statistical framework for single cell and bulk RNA-sequencing data. *Ann. Appl. Stat.*, **12**, 609–632.
32. Danziger,S.A., Gibbs,D.L., Shmulevich,I., McConnell,M., Trotter,M.W.B., Schmitz,F., Reiss,D.J. and Ratushny,A.V. (2019) ADAPTS: Automated deconvolution augmentation of profiles for tissue specific cells. *PLoS One*, **14**, e0224693.
33. Royston,J.P. (1982) An extension of Shapiro and Wilk's W test for normality to large samples. *Appl. Stat.*, **31**, 115.
34. Frishberg,A., Peshes-Yaloz,N., Cohn,O., Rosentul,D., Steuerman,Y., Valadarsky,L., Yankovitz,G., Mandelboim,M., Iraqi,F.A., Amit,I. *et al.* (2019) Cell composition analysis of bulk genomics using single-cell data. *Nat. Methods*, **16**, 327–332.
35. Aran,D., Hu,Z. and Butte,A.J. (2017) xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.*, **18**, 220.
36. Wickham,H. (2011) ggplot2. Wiley Interdiscip. *Rev. Comput. Mol. Sci.*, **3**, 180–185.
37. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
38. Reimand,J., Arak,T., Adler,P., Kolberg,L., Reisberg,S., Peterson,H. and Vilo,J. (2016) g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.*, **44**, W83–W89.
39. Reimand,J., Isserlin,R., Voisin,V., Kucera,M., Tannus-Lopes,C., Rostamianfar,A., Wadi,L., Meyer,M., Wong,J., Xu,C. *et al.* (2019) Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.*, **14**, 482–517.
40. Wold,S., Esbensen,K. and Geladi,P. (1987) Principal component analysis. *Chemom. Intell. Lab. Syst.*, **2**, 37–52.
41. Frazee,A.C., Jaffe,A.E., Langmead,B. and Leek,J.T. (2015) Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, **31**, 2778–2784.
42. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
43. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
44. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
45. Zhang,Y., Parmigiani,G. and Johnson,W.E. (2020) ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom. Bioinform.*, **2**, lqaa078.
46. Leek,J.T., Johnson,W.E., Parker,H.S., Jaffe,A.E. and Storey,J.D. (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**, 882–883.
47. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.
48. Picelli,S., Faridani,O.R., Björklund,A.K., Winberg,G., Sagasser,S. and Sandberg,R. (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.*, **9**, 171–181.
49. Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.
50. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.
51. Ilicic,T., Kim,J.K., Kolodziejczyk,A.A., Bagger,F.O., McCarthy,D.J., Marioni,J.C. and Teichmann,S.A. (2016) Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.*, **17**, 29.
52. Smedley,D., Haider,S., Durinck,S., Pandini,L., Provero,P., Allen,J., Arnaiz,O., Awedh,M.H., Baldock,R., Barbiera,G. *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.*, **43**, W589–W598.
53. Hänzelmann,S., Castelo,R. and Guinney,J. (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.
54. Diaz-Mejia,J.J., Meng,E.C., Pico,A.R., MacParland,S.A., Ketela,T., Pugh,T.J., Bader,G.D. and Morris,J.H. (2019) Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data. *F1000Res.*, **8**, 296.
55. Mehta,C.R. and Patel,N.R. (1983) A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *J. Am. Stat. Assoc.*, **78**, 427–434.
56. Leinonen,R., Sugawara,H., Shumway,M. and International Nucleotide Sequence Database Collaboration (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–21.
57. Wickham,H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer, NY.

58. Hafemeister,C. and Satija,R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome biol.*, **20**, 296.
59. Finak,G., McDavid,A., Yajima,M., Deng,J., Gersuk,V., Shalek,A.K., Slichter,C.K., Miller,H.W., McElrath,M.J., Prlic,M. *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.
60. McDavid,A., Finak,G., Chattopadyay,P.K., Dominguez,M., Lamoreaux,L., Ma,S.S., Roederer,M. and Gottardo,R. (2013) Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics*, **29**, 461–467.
61. Newman,A.M., Liu,C.L., Green,M.R., Gentles,A.J., Feng,W., Xu,Y., Hoang,C.D., Diehn,M. and Alizadeh,A.A. (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, **12**, 453–457.
62. Sato,Y. and Yanagita,M. (2017) Resident fibroblasts in the kidney: a major driver of fibrosis and inflammation. *Inflamm. Regen.*, **37**, 17.
63. Verma,S.K. and Molitoris,B.A. (2015) Renal endothelial injury and microvascular dysfunction in acute kidney injury. *Semin. Nephrol.*, **35**, 96–107.
64. Havasi,A. and Dong,Z. (2016) Autophagy and tubular cell death in the kidney. *Semin. Nephrol.*, **36**, 174–188.
65. Monroy,M.A., Fang,J., Li,S., Ferrer,L., Birkenbach,M.P., Lee,I.J., Wang,H., Yang,X.-F. and Choi,E.T. (2015) Chronic kidney disease alters vascular smooth muscle cell phenotype. *Front. Biosci. (Landmark Ed)*, **20**, 784–795.
66. Karaiskos,N., Rahmatollahi,M., Boltengagen,A., Liu,H., Hoehne,M., Rinschen,M., Schermer,B., Benzing,T., Rajewsky,N., Kocks,C. *et al.* (2018) A single-cell transcriptome atlas of the mouse glomerulus. *J. Am. Soc. Nephrol.*, **29**, 2060–2068.
67. Ascon,D.B., Ascon,M., Satpute,S., Lopez-Briones,S., Racusen,L., Colvin,R.B., Soloski,M.J. and Rabb,H. (2008) Normal mouse kidneys contain activated and CD3+CD4- CD8- double-negative T lymphocytes with a distinct TCR repertoire. *J. Leukoc. Biol.*, **84**, 1400–1409.
68. Kim,K.W., Kim,B.-M., Doh,K.C., Cho,M.-L., Yang,C.W. and Chung,B.H. (2018) Clinical significance of CCR7+CD8+ T cells in kidney transplant recipients with allograft rejection. *Sci. Rep.*, **8**, 8827.
69. Braga,T.T., Correa-Costa,M., Silva,R.C., Cruz,M.C., Hiyane,M.I., da Silva,J.S., Perez,K.R., Cuccovia,I.M. and Camara,N.O.S. (2018) CCR2 contributes to the recruitment of monocytes and leads to kidney inflammation and fibrosis development. *Inflammopharmacology*, **26**, 403–411.
70. Park,J.-G., Lee,C.-R., Kim,M.-G., Kim,G., Shin,H.M., Jeon,Y.-H., Yang,S.H., Kim,D.K., Joo,K.W., Choi,E.Y. *et al.* (2020) Kidney residency of VISTA-positive macrophages accelerates repair from ischemic injury. *Kidney Int.*, **97**, 980–994.
71. Bonner,W.A., Hulett,H.R., Sweet,R.G. and Herzenberg,L.A. (1972) Fluorescence activated cell sorting. *Rev. Sci. Instrum.*, **43**, 404–409.
72. DiSpirito,J.R., Zemmour,D., Ramanan,D., Cho,J., Zilionis,R., Klein,A.M., Benoist,C. and Mathis,D. (2018) Molecular diversification of regulatory T cells in nonlymphoid tissues. *Sci. Immunol.*, **3**, eaat5861.
73. Schmidt,D., Wilson,M.D., Spyrou,C., Brown,G.D., Hadfield,J. and Odom,D.T. (2009) ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods*, **48**, 240–248.
74. Kuleshov,M.V., Jones,M.R., Rouillard,A.D., Fernandez,N.F., Duan,Q., Wang,Z., Koplev,S., Jenkins,S.L., Jagodnik,K.M., Lachmann,A. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90–W97.
75. Vieth,B., Parekh,S., Ziegenhain,C., Enard,W. and Hellmann,I. (2019) A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.*, **10**, 4667.
76. Cusanovich,D.A., Hill,A.J., Aghamirzaie,D., Daza,R.M., Pliner,H.A., Berletch,J.B., Filippova,G.N., Huang,X., Christiansen,L., DeWitt,W.S. *et al.* (2018) A single-cell atlas of *in vivo* mammalian chromatin accessibility. *Cell*, **174**, 1309–1324.
77. Buenrostro,J.D., Wu,B., Litzenburger,U.M., Ruff,D., Gonzales,M.L., Snyder,M.P., Chang,H.Y. and Greenleaf,W.J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**, 486–490.
78. Hui,T., Cao,Q., Wegrzyn-Woltosz,J., O'Neill,K., Hammond,C.A., Knapp,D.J.H.F., Laks,E., Moksa,M., Aparicio,S., Eaves,C.J. *et al.* (2018) high-resolution single-cell DNA methylation measurements reveal epigenetically distinct hematopoietic stem cell subpopulations. *Stem Cell Rep.*, **11**, 578–592.
79. Karemaker,I.D. and Vermeulen,M. (2018) Single-cell DNA methylation profiling: technologies and biological applications. *Trends Biotechnol.*, **36**, 952–965.