

METHOD

Open Access



# treekoR: identifying cellular-to-phenotype associations by elucidating hierarchical relationships in high-dimensional cytometry data

Adam Chan<sup>1,2</sup>, Wei Jiang<sup>3,4</sup>, Emily Blyth<sup>3,4,5</sup>, Jean Yang<sup>1,2,6</sup> and Ellis Patrick<sup>1,3,6\*</sup> 

\* Correspondence: [ellis.patrick@sydney.edu.au](mailto:ellis.patrick@sydney.edu.au)

<sup>1</sup>School of Mathematics and Statistics, The University of Sydney, Sydney, New South Wales, Australia  
<sup>3</sup>Centre for Cancer Research, Westmead Institute for Medical Research, The University of Sydney, Sydney, New South Wales, Australia  
Full list of author information is available at the end of the article

## Abstract

High-throughput single-cell technologies hold the promise of discovering novel cellular relationships with disease. However, analytical workflows constructed for these technologies to associate cell proportions with disease often employ unsupervised clustering techniques that overlook the valuable hierarchical structures that have been used to define cell types. We present treekoR, a framework that empirically recapitulates these structures, facilitating multiple quantifications and comparisons of cell type proportions. Our results from twelve case studies reinforce the importance of quantifying proportions relative to parent populations in the analyses of cytometry data — as failing to do so can lead to missing important biological insights.

## Introduction

High-parameter cytometry assays have provided biomedical scientists with an unprecedented detail of the cellular heterogeneity of patient samples. Flow and mass cytometers are able to characterize cells by measuring up to fifty extracellular antigens [1], with single-cell sequencing platforms able to measure thousands of intracellular RNA molecules [2]. Unfortunately, this ground-breaking capacity to characterize cells to this depth has provided a computational challenge for bioinformaticians to efficiently glean meaningful information from the deluge of single-cell data. Given that most novel analytical methods neglect the hierarchical relationships in single-cell data, there exists an opportunity to use these relationships to identify robust and interpretable associations between cell subsets and patient clinical end points or ex vivo interventions.

To compare the abundance of cell subsets between samples, there has been a decades-long legacy of either quantifying a cell type as the proportion relative to all cells in a sample (*%total*), or, as the proportion relative to a parent population



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

(%parent) [3–5]. The latter of these quantifications is derived naturally from the way that cell subsets have traditionally been annotated via a process called sequential manual gating [6]—where 2D scatter plots are drawn using certain markers and gated to identify cell populations in a sequential manner. For example, regulatory T cells (Tregs) could be identified by first gating out CD3+ and CD4+ cells to identify CD4+ T cells and then further gating on CD25lo and CD127+ to isolate the CD4+ Tregs [7]. This gating strategy naturally lends itself to quantification of cell types relative to their parent lymphocyte populations. These quantifications are robust to changes in unrelated subsets. The main drawbacks of this method however are its reliance on time-consuming manual gating, which has become impractical for high-parameter assays [8], and the substantial reliance on expert knowledge which may bias analysis towards known and expected relationships.

As an alternative cell type identification strategy to manual gating, unsupervised clustering of cells has been used to circumvent the challenges of sequentially gating high-dimensional cytometry data. These automated methods are able to stratify cell subsets without necessarily having a predetermined hypothesis or sequential gating strategy. Many methods, including SPADE [9], Citrus [10], FlowSOM [11], Phenograph [12], SC3 [13], and scClust [14] have been utilized frequently in the analysis of high-dimensional cytometry data to identify cell populations. While they have significantly improved the efficiency in which scientists can analyze these datasets, typical analyses employing these methods only explore the changes in cell types as a %total, neglecting the complex hierarchical proportions inherent in single-cell data. In other words, these methods fail to measure cell types as a %parent, which cytometry analysts have traditionally used in manual gating workflows.

A number of unsupervised clustering methods and data-driven workflows have been developed to explore the hierarchical nature of cytometry data. SPADE and FlowSOM utilize minimal spanning trees over clustering as a visualization tool. Citrus employs hierarchical clustering and regularized supervised learning algorithms to identify stratifying populations of cells on each level of aggregation. The method treeclimbR [15] aims to pinpoint an ideal resolution of cell populations via a hierarchical tree. Although these methods acknowledge the importance of visualizing the hierarchical aspect of single-cell cytometry data, they do not typically incorporate such information in their association analysis. That is, they do not by default quantify the abundance of cell types as a %parent and test if these compositions are associated with a treatment or phenotype of interest.

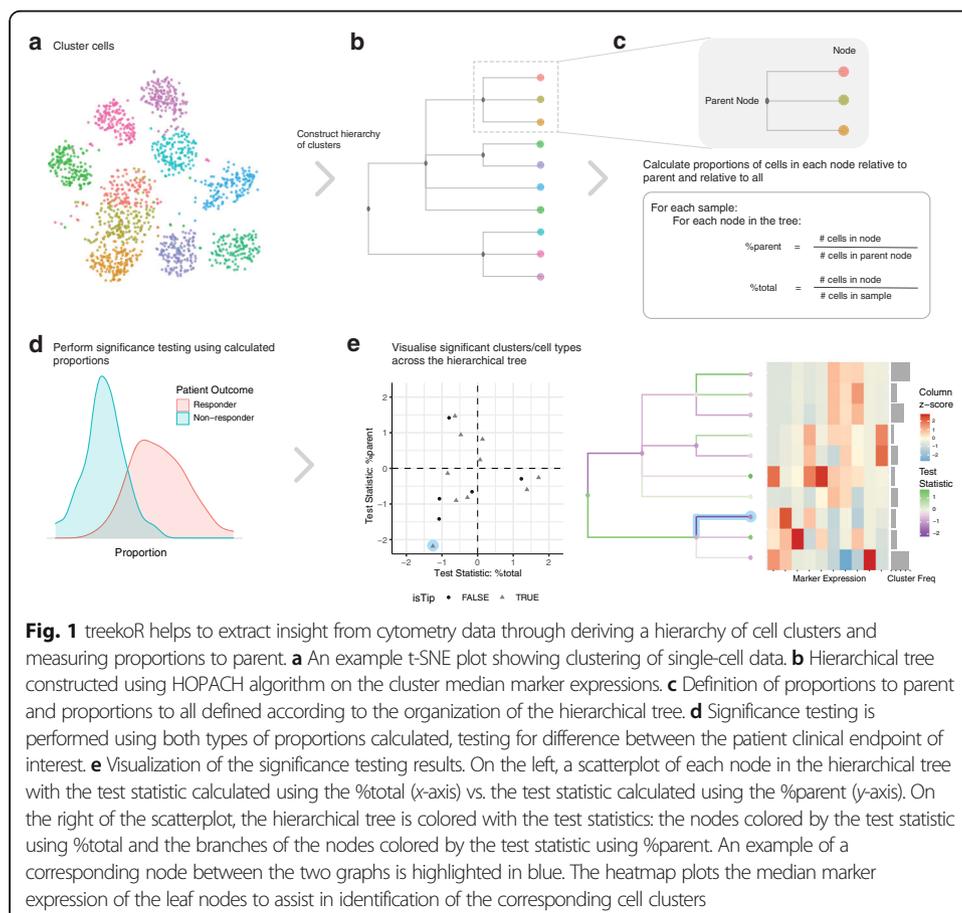
To this end, we have developed treekoR, a novel framework that makes use of cell type identification from unsupervised clustering techniques while acknowledging the hierarchical nature of single-cell cytometry data to discover robust and interpretable associations between cell subsets and patient outcomes. TreekoR achieves this by (1) algorithmically deriving the hierarchy of cell type clusters, followed by (2) incorporating this hierarchical information via measuring the %parent for each cell type. These derived proportions can then be used in significance testing and classification models to determine associations with clinical outcomes. Further to this, treekoR provides a general framework that is flexible to the clustering approach, hierarchical aggregation method, and type of significance testing used. This framework allows analysts to

generate insight from the complex hierarchical relationships present in single-cell cytometry data, which are often overlooked with existing automated clustering methods.

## Results

### treekoR algorithmically derives cell type hierarchies to quantify %parent

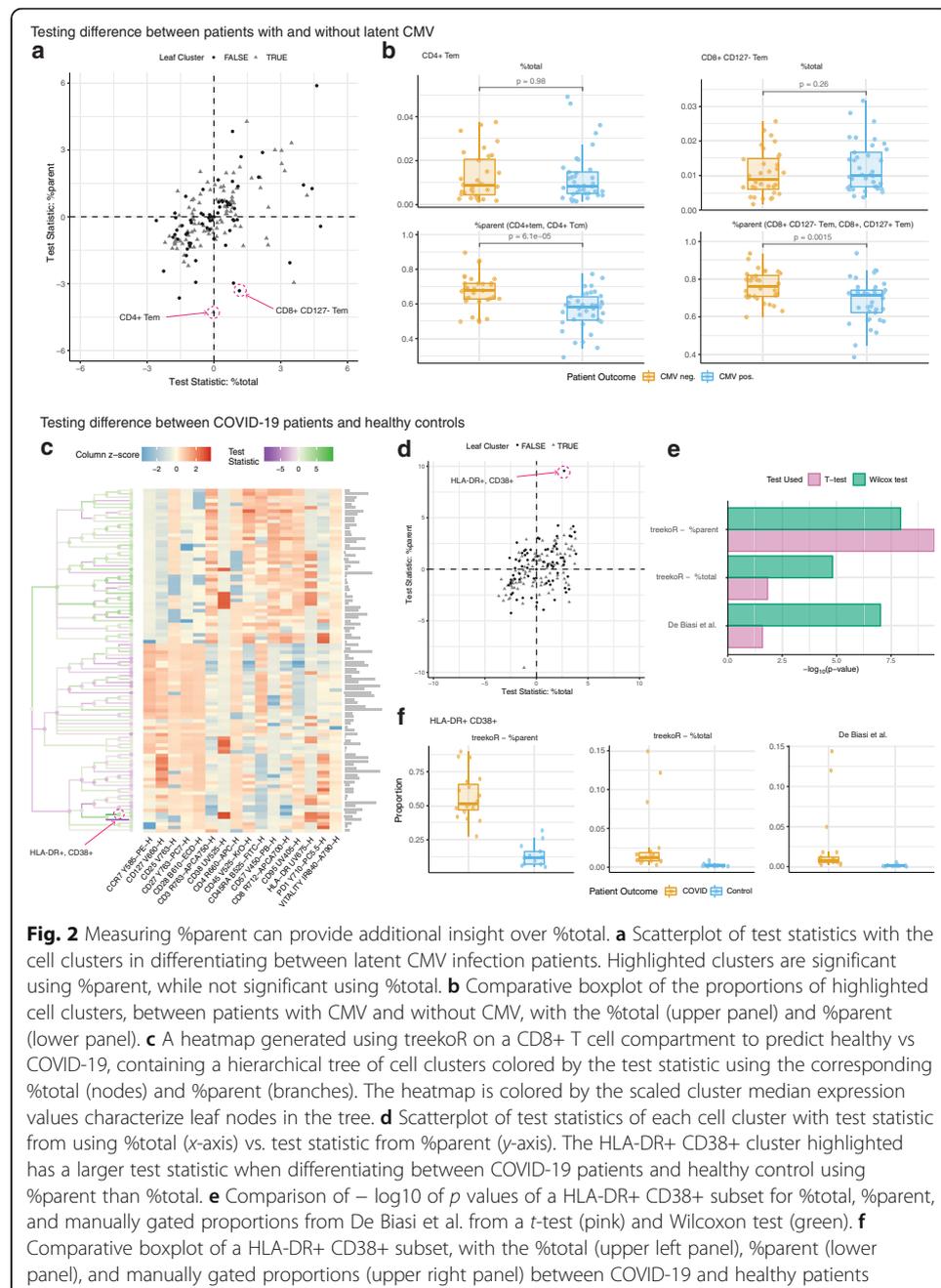
We present treekoR, an analytical framework that recognizes and incorporates the hierarchical relationships inherent in cytometry data. The treekoR package is implemented in R and uses an automated workflow to identify cellular associations with a patient outcome through five steps (Fig. 1): (1) cluster the data using an automated method, (2) aggregate clusters into a tree using a hierarchical clustering algorithm, (3) calculate the %total (the proportion of a cell type relative to all cells in a sample) and %parent (the proportion of a cell type relative to a parent population of cells, in this case the cells in the parent node) of cells in each node in the tree, (4) perform significance tests using both of these proportions against a clinical end point, and (5) visualize the significance results on the tree. The %parent calculated by treekoR aims to emulate the proportions naturally derived when using sequential manual gating, which are not typically calculated in workflows exclusively using unsupervised clustering methods. Our comparative procedure uncovers important associations with a clinical end point of interest



by visualizing both quantifications of cell type proportions derived from the data. Further details are provided under “Methods.”

**treekoR generates biological insight exclusive to %parent in example cytometry datasets**

We illustrate the ability of treekoR to generate additional biological insight by applying the framework to a CyTOF study of latent Cytomegalovirus (CMV) [16]. After clustering cells into one hundred cell subsets, quantifying the %total and %parent for each, and testing for associations between CMV positivity and %total or %parent (Fig. 2a); we observed a reduction in CD4+ Tem cells in CMV-positive patients using %parent ( $p =$



**Fig. 2** Measuring %parent can provide additional insight over %total. **a** Scatterplot of test statistics with the cell clusters in differentiating between latent CMV infection patients. Highlighted clusters are significant using %parent, while not significant using %total. **b** Comparative boxplot of the proportions of highlighted cell clusters, between patients with CMV and without CMV, with the %total (upper panel) and %parent (lower panel). **c** A heatmap generated using treekoR on a CD8+ T cell compartment to predict healthy vs COVID-19, containing a hierarchical tree of cell clusters colored by the test statistic using the corresponding %total (nodes) and %parent (branches). The heatmap is colored by the scaled cluster median expression values characterize leaf nodes in the tree. **d** Scatterplot of test statistics of each cell cluster with test statistic from using %total (x-axis) vs. test statistic from %parent (y-axis). The HLA-DR+ CD38+ cluster highlighted has a larger test statistic when differentiating between COVID-19 patients and healthy control using %parent than %total. **e** Comparison of  $-\log_{10}$  of  $p$  values of a HLA-DR+ CD38+ subset for %total, %parent, and manually gated proportions from De Biasi et al. from a  $t$ -test (pink) and Wilcoxon test (green). **f** Comparative boxplot of a HLA-DR+ CD38+ subset, with the %total (upper left panel), %parent (lower panel), and manually gated proportions (upper right panel) between COVID-19 and healthy patients

$6.1 \times 10^{-5}$ , FDR =  $3.33 \times 10^{-3}$ ), yet no association was observed using %total ( $p = 0.9$ , FDR = 0.99). The higher proportion of CD4 + Tem relative to its parent cluster (CD4 + Tem and CD4 + Tcm) in CMV-negative patients as compared to CMV-positive patients is in keeping with known effector memory cell function in cytokine secretion and viral clearance. Similarly, observed a nominally significant negative association between CMV positivity and CD8+ CD127- Tem cells using %parent ( $p = 1.5 \times 10^{-3}$ , FDR =  $3.5 \times 10^{-2}$ ), but not with %total ( $p = 0.26$ , FDR = 0.69) (Fig. 2b). This lower proportion of CD8+ CD127- Tem cells relative to its parent (CD8+ CD127- and CD8+ CD127+ Tem) in CMV-positive patients as compared with CMV-negative patients suggests a role for differential CD127 expression in chronic/persistent infection. Together, this suggests that if the %parent of these cell types had not been measured, we would have been unable to discover the cellular relationships between CD4+ Tem and CD8+ CD127- Tem with CMV infection.

When applied to a flow cytometry dataset profiling CD8+ T cells in COVID-19 patients and healthy controls [4], treekoR highlighted a highly activated HLA-DR+ CD38+ CD8+ T cell subset whose %parent provided a more robust association with COVID-19 response than its %total. After applying FlowSOM to cluster cell types (Fig. 2c), we discovered a HLA-DR+ CD38+ CD8+ T cell whose %parent is greater in COVID-19 patients than healthy controls ( $p = 3.19 \times 10^{-10}$ , FDR =  $2.76 \times 10^{-8}$ ) (Fig. 2d). This strong association is observed regardless of whether the %parent is modelled as continuous or count data (Additional file 1: Figure S1). However, this population only appeared marginally associated with COVID-19 response using %total ( $p = 1.49 \times 10^{-2}$ , FDR =  $7.6 \times 10^{-2}$ ). In contrast, De Biasi et al. had reported a manually gated HLA-DR+ CD38+ CD8+ T cell population changing when using %total ( $p = 9.70 \times 10^{-8}$ ). The difference in conclusion between using %total from FlowSOM and the manually gated population from De Biasi et al. is solely attributed to our use of a *t*-test and De Biasi et al.'s use of the Wilcoxon rank sum test (Fig. 2e), which is robust to the outliers observed in the %total quantification (Fig. 2f). When a Wilcoxon rank sum test is used on our %total ( $p = 1.55 \times 10^{-5}$ , FDR =  $1.12 \times 10^{-3}$ ) and %parent ( $p = 1.15 \times 10^{-8}$ , FDR =  $9.96 \times 10^{-7}$ ), the association is also observed, but not observed when a *t*-test is used on De Biasi et al.'s manually gated population ( $p = 2.57 \times 10^{-2}$ ). The presence of this association in treekoR's %parent regardless of the significance test used illustrates that quantifying the proportion of HLA-DR+ CD38+ to a parent population (HLA-DR+ CD38+ and HLA-DR+ CD38-) can adjust for large fluctuations in cell type compositions and allow subtle changes in proportion to be robustly quantified. Across both the COVID-19 and CMV case studies, we highlight two perspectives of cell type proportions, %total and %parent, which offer biological information that may be potentially missed if only one was measured.

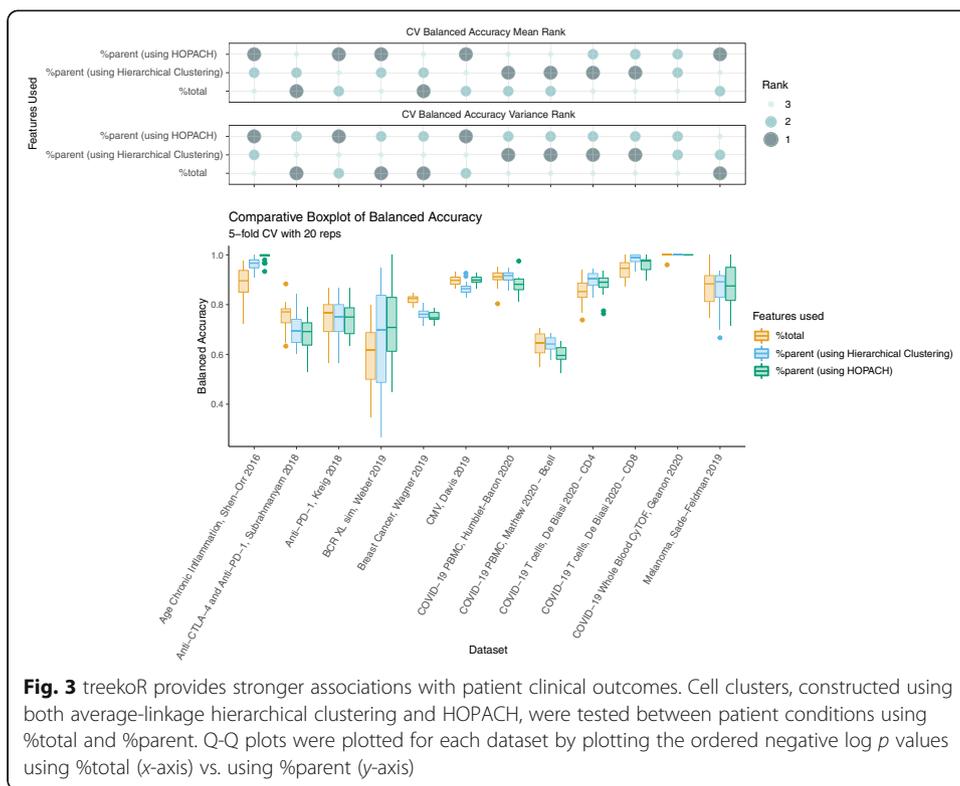
### **The %parent of cell types yields strong associations with clinical outcomes across several datasets in our benchmark**

We observed a greater discrimination between binary outcomes through quantifying proportions as %parent than %total in several datasets. We compared twelve case studies consisting of seven CyTOF datasets, four flow cytometry datasets, and a single-cell RNA sequencing (scRNA-seq) dataset (Table 1). Further, we also used two hierarchical

**Table 1** Benchmark datasets. Eleven published datasets were used to compare %total and %parent in significance testing and classification using the treekoR workflow. "Name" is used to refer to each dataset throughout the manuscript

Name	Technology	Description	Number of cells	Number of samples	Outcome or response variable	References
Age chronic	CytoF	Age chronic inflammation predicting young vs old	1036209	29	Young / old	Shen-Orr et al. 2016 [18] Immport [30] SDY887 dataset
Anti-CTLA-4 and anti-PD-1	CytoF	Predicting response vs non-response in anti-CTLA-4 and anti-PD-1 treatments	7264780	24	Response / non-response to treatment	Subrahmanyam et al. 2018 [21]
Anti-PD-1	CytoF	Predicting response vs non-response in anti-PD-1 treatment	85718	20	Response / non-response to treatment	Kreig et al. 2018 [31]
BCR-XL-sim	CytoF	Detecting samples with stimulated B cells	88435	16	Spiked / non-spiked	Weber et al. 2019 [23]
Breast cancer tumor	CytoF	Predicting tumor in breast cancer samples	855914	194	Tumor/ non-tumor breast cancer samples	Wagner et al. 2019 [32]
CMV	CytoF	Predicting positive vs negative CMV titer results in influenza patients	18153877	69	Positive/ negative results from CMV titer	Tomic et al. 2019 [16] Immport [30] SDY478 dataset
COVID-19 whole blood CyTOF	CytoF	Profiling whole blood to predict COVID-19 vs. healthy patients	4747543	21	COVID-19 / healthy control	Geanon et al. 2021 [33]
COVID-19 PBMCs	Flow cytometry	Predicting between ICU vs. hospital ward COVID-19 patients	4790053	38	ICU / ward	Humblet-Baron et al. 2021 [34]
COVID-19 PBMC CD8+ non-naive T cells	Flow cytometry	Profile of CD8+ Non-Naive T Cells to distinguish recovered from COVID-19 vs. healthy	11591741 (60% of cells were sampled and analyzed)	168	COVID-19 recovered / healthy	Mathew et al. 2020 [35]
COVID-19 T cells	Flow cytometry	T cell compartment samples (CD4 and CD8) to predict healthy vs COVID-19	5000	31	COVID-19 / healthy control	De Biasi et al. 2020 [4]
Melanoma	scRNA-seq	Predicting response to checkpoint immunotherapy in melanoma	5928	19	Responder/ non-responder	Sade-Feldman et al. 2019 [36]

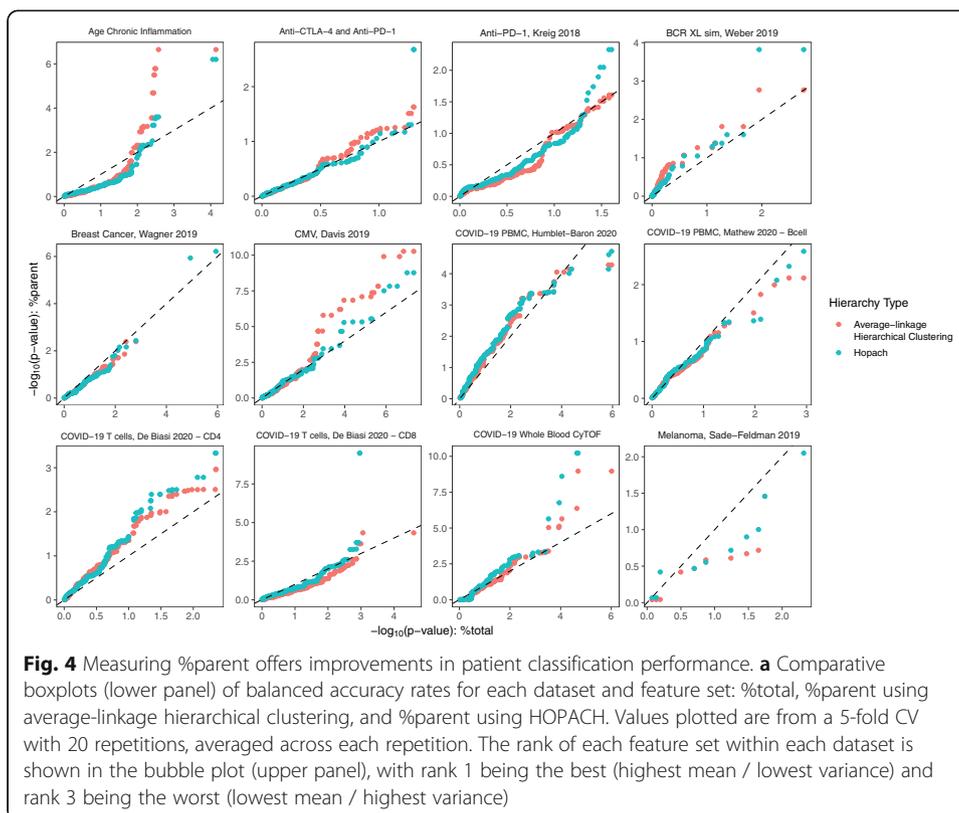
clustering algorithms, HOPACH [17] and average-linkage hierarchical clustering, with both generating different estimates of %parent (Additional file 1: Figure S2). After testing for differences in cell type proportions between the patient conditions, we compared the ordered negative log  $p$  values of each cell population from using %total against the ordered negative log  $p$  values from using %parent (Fig. 3). Across all twelve case studies, we were able to determine whether performing significance testing using %parent provided comparatively stronger associations with the patient outcome than %total—evident in instances where points conspicuously lay above the dashed identity



line. Across half of the investigated datasets, in particular CMV [16] and Age Chronic [18], the cell type proportion with highest significance was obtained from measuring its %parent. Further to this, the choice of hierarchical aggregation techniques produced variations in clinical association, suggesting that using different cell type trees can help analysts uncover a wider scope of associations. The benchmark exemplifies the importance of measuring both %parent and %total so as not to miss pertinent clinical associations.

**Multivariate classification of clinical outcomes in cytometry data can be improved by measuring %parent**

High-dimensional single-cell data have been used to construct models to classify patients to help scientists discover and understand associations with a clinical outcome [19–22]. We evaluated classification performance using either %total or %parent as feature sets in several datasets with binary outcomes (e.g., responder vs. non-responder, COVID-19 vs. healthy control), to determine that the incorporation of %parent features in multivariate classifications models can help improve patient classification. There were various differences in balanced accuracy between using %total and %parent (using either HOPACH or hierarchical clustering with average linkage) in each dataset (Fig. 4). The datasets with the biggest increase in balanced accuracy by using %parent were the BCR-XL-sim data [23] and Age Chronic data [18]. In the BCR-XL-sim semi-simulated dataset, we predicted which samples contained stimulated B cells. Using only %total as features produced a mean balanced accuracy of 59%, compared to 73% using %parent derived from HOPACH. In the Age Chronic CyTOF dataset, classifiers were constructed to discriminate between older and younger adults using their immune response signatures to influenza vaccination. Here, we show that using %parent (99%) also gives



a higher mean balanced accuracy than %total (88%). These results support the notion that failing to measure %parent can sometimes mean neglecting important signals when trying to predict a patient’s clinical outcome in high-dimensional cytometry datasets.

Quantifying multiple views of cell type proportions can provide greater insight into single-cell cytometry data and patient clinical outcomes. In our classification benchmark, we compared the use of %total, %parent (using hierarchical clustering), and %parent (using HOPACH) cell type proportions. Exploring hierarchical representations via treekoR can help to elucidate a broader scope of %parent relationships that exist within cytometry data (Additional file 1: Figure S2). When each feature set was ranked using the mean and standard deviation of the balanced accuracy in each dataset (Fig. 4), no single quantification of proportion performed the best for prediction of patient outcomes across all analyzed cytometry datasets. The differences in rank however mean that each type of proportion quantification provided a different perspective of the data. Depending on the dataset, one approach may provide a greater coverage of the signal present within the data through a higher balanced accuracy. This further supports the idea that proportions measured as %total should not be the only proportions measured in cytometry analysis workflows, particularly when searching for the most predictive features in distinguishing between patient clinical outcomes and understanding the complex relationships that exist. It is therefore imperative that proportions are quantified as both %parent and %total for the effective analysis of cytometry data, as it offers more thorough examination of this data.

## Discussion

In this paper, we examined several high-dimensional single-cell datasets to demonstrate the importance of measuring both %parent and %total proportions, the use of %parent for classification, and the consequences of using different hierarchical aggregation techniques to empirically derive cell type proportions. Overall, we accentuated the importance of analyzing high-dimensional cytometry data using ideas from both traditional manual gating and unsupervised clustering techniques and provide a general framework, *treekoR*, which allows analysts to do so while overcoming key pitfalls of both approaches.

The *treekoR* framework allows scientists to select their own clustering algorithm for determination of cell types and hierarchical aggregation technique for the construction of cell type trees. While there have been numerous comparisons of clustering methods of cytometry data [19, 24–27], there have not been as many comparisons of hierarchy construction techniques in the context of cell type hierarchies [9, 11]. We show through the use of HOPACH and average-linkage hierarchical clustering that the choice of hierarchical aggregation technique can have noteworthy effects on downstream analysis, and suggest multiple other techniques that could also be used to produce distinct cell type trees. However, no formal evaluation to determine the most “suitable” technique was performed throughout our analyses. Since scientists have unique and personal workflows for hierarchically analyzing cell types, there is significant room to explore what an appropriate cell type hierarchy might entail and determine a corresponding standard or measure which scientists can use to evaluate this. The definition for the most “suitable” hierarchical aggregation technique, whether it is the technique which produces the most interpretable hierarchy or produces the %parent proportions most associated with a clinical outcome, has yet to be elaborated.

In *treekoR*, we defined %parent as the proportion of a cell type relative to its direct parent in the cell type hierarchy. This proportion could be calculated using a broader parent (e.g., a higher ancestor) cell type in the hierarchy, which could lead to either a more interpretable and familiar cell type %parent or reduce the burden of multiple hypothesis testing. Since the scope of proportions to be calculated becomes much larger when numerous measurements of %parent for a single-cell type are allowed, there exists a challenge in determining which %parent to calculate, particularly as the number of hypothesis tests increases. We do not currently address either of these points in our workflow. To overcome this challenge, a standard set of reference cell types can be determined to calculate %parent from. These reference cell types could be deduced in a semi-supervised fashion where analysts manually select them, or in a completely unsupervised manner by using a data-driven method (such as *treeclimbR* [15]). This would limit the number of proportions calculated and potentially provide more biologically relevant %parent. Another approach to this issue could be implementing a multiple hypothesis correction that caters for the hierarchical nature of these proportions.

Care is required in the comparison of statistical significance between the %total and %parent of a cell type. The derived  $p$  values from significance testing inherently come from two distinct statistical hypotheses. Therefore, the user should not conclude that one proportion is a better metric based solely on its  $p$  value, or say that one proportion is more relevant than the other. Rather the %total and %parent provide two complementary views, both of which may be objective and biologically relevant. Depending on

the datasets, one quantification of cell type proportions may provide a stronger association with a clinical outcome of interest, this nuance is important to note.

In summary, we present a framework that is general in nature, allowing scientists to choose algorithms appropriate to their dataset to glean more information than typical analyses. It is our broader intention to emphasize the importance of measuring %parent in the analysis of cytometry data—and that these hierarchical proportions should not be overlooked as researchers move towards more efficient and automated approaches of analysis. As high-dimensional cytometry data become more ubiquitous in helping scientists understand the underlying biological process behind patient diseases, such as influenza and COVID-19, we envision that the implementation of treekoR will assist in unravelling the cell type heterogeneity present in these complex patient diseases.

## Methods

### Overview of treekoR

treekoR is performed in five steps: (i) cluster the data using an automated method, (ii) aggregate clusters into a tree using a hierarchical clustering algorithm, (iii) calculate the %total and %parent of cells in each node of the tree, (iv) perform significance tests using both of these proportions against a clinical end point, and (v) visualize the significance results on the tree. The steps are described in detail below, along with the parameters used in the analyses throughout the manuscript.

- (i) *Clustering*. Unsupervised clustering was performed using the FlowSOM [11] algorithm as part of the CATALYST [28] package in R [29], using a  $10 \times 10$  grid. Cells are over-clustered to try to account for all cell types present within the data and to avoid missing rare cell populations (any superfluous clusters are then naturally aggregated in the hierarchical clustering step). For the datasets that were provided with previously analyzed or manually gated cell types, those cell types were used instead of the FlowSOM clustering.
- (ii) *Construction of hierarchy*. Following clustering of the data, the scaled median marker expression for each cluster was calculated and used to construct a hierarchical tree. Several hierarchical clustering techniques can be used in treekoR and are included in the R stats hclust function [29]. These include HOPACH and agglomerative hierarchical clustering using average linkage, Ward linkage, single linkage, complete linkage, and McQuitty. HOPACH allows for multiple children per node while other included methods only cater for two children per node. Throughout the analysis, we used two main methods for hierarchical aggregation: HOPACH (with  $K = 5$  maximum children per parent node) and average-linkage hierarchical clustering.
- (iii) *Calculation of proportions*. After clustering, when a hierarchical tree of cell types has been established in the data, the proportions of these clusters are quantified. For each patient, the proportions of cells belonging to the clusters in each node of the tree are measured relative to their total number of the cells, referred to as %total. In addition, the clusters in each node of the tree are measured as a proportion of the cluster in the direct parent node of the tree, referred to as %parent.

- (iv) *Significance testing.* For each node in the hierarchical tree on the clusters, significance testing is then performed using a two-sample  $t$ -test for equal means between the desired patient outcome using both the %total and %parent.
- (v) *Visualization.* The results of these proportions can be then visualized through a colored tree plotted next to a corresponding heatmap. The heatmap displays the median scaled marker expressions of each cluster to help understand what cell type each cluster may represent, and the tree not only reveals how clusters have been hierarchically aggregated, but is colored on each node by the test statistic obtained when testing using %total of that node, with the branch connecting the child to the parent colored by the test statistic obtained when testing using the %parent of the child node.

### Benchmark data and data processing

The twelve benchmarking datasets consist of seven CyTOF, four flow cytometry, and one single-cell RNA-seq dataset as shown in Table 1. In the flow cytometry datasets, COVID-19 T cells were counted as two datasets—CD4 and CD8 T cells.

### Data normalization

For each of the cytometry datasets, we applied an arcsinh transformation with a co-factor of 5 on the expression values. The samples were then filtered to only include the patients with the clinical end points of interest. For analysis of the CMV dataset, 66.67% of cells were randomly subsampled and gated for live intact cells before transforming.

### Calculation of proportions

For each of the patients/samples, the proportions of each of the FlowSOM clusters or cell types were calculated as %total, as well as %parent from a HOPACH [17] tree and an average-linkage hierarchical clustering tree. The %parent for each cluster in each sample is calculated as the (# cells in a cluster) / (# cells in a cluster + # cells in sibling clusters). The %total is calculate as (# cells in a cluster) / (# cells in sample).

### Hypothesis testing

For each of the cell types/clusters, a two-sample  $t$ -test was used to test if there was a significant difference in mean proportion between the binary clinical outcome of interest, using both %total and %parent. In our COVID-19 T cells and CMV case studies,  $p$  value adjustment was performed using the FDR method, while  $p$  value adjustment was not performed in the benchmark comparison.

The  $p$  values obtained from using  $t$ -tests on %parent and %total were compared to the  $p$  values obtained from the count models: edgeR [37] and generalized linear mixed models [38] (GLMM). The method of testing the difference between patient conditions using these count models was adapted from differential abundance testing in the diffcyt [23] package. These methods were naturally able to test differences in %total; however, they had to be adapted slightly to test for %parent. More specifically, we tested %parent in edgeR by using the counts of each cluster, using library sizes of 1, and using the number of cells in the parent clusters as an offset. We tested %parent using GLMM by

treating %parent as the dependent variable, condition as an independent variable with fixed effect, sample as an independent variable with random effect, and using the number of cells in the parent clusters as weights.

### Classification

The %total and %parent proportions were then used as features separately, for sake of comparison, to predict the binary patient clinical end point. For each feature set and dataset combination, we trained a random forest (using mlr3 [39]) with 500 trees in each iteration of a 5-fold cross validation with 20 repetitions. The balanced accuracy was measured in each iteration of the cross validation and used to compare predictive power between the feature sets.

### Supplementary information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-021-02526-5>.

**Additional file 1.** Supplementary materials [42, 43].

**Additional file 2.** Review history.

### Acknowledgements

The authors thank all their colleagues, particularly at The University of Sydney, School of Mathematics and Statistics, Charles Perkins Center, for their support and intellectual engagement.

### Review history

The review history is available as Additional file 2.

### Peer review information

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

### Authors' contributions

EP conceived and designed the study with input from JY. EP and AC led the treekoR method development and JY developed and guided the evaluation data analysis. AC curated the benchmarking data, implemented all data analytics, and developed the R package with guidance from EP. WJ and EB contributed the biological interpretation of statistical findings. All authors wrote, read, reviewed the manuscript and approved the final version.

### Funding

The following sources of funding for each author, and for the manuscript preparation, are gratefully acknowledged: Australian Research Council Discovery Project grant (DP170100654) to JYHY; Australian Research Council Discovery Early Career Researcher Award (DE200100944) funded by the Australian Government to EP; Australian Government Research Training Program (RTP) Scholarship to AC; PhD scholarship from the Haematology Society of Australia and New Zealand (HSANZ) and Leukaemia Foundation Australia to WJ. EB is a NSW Cancer Institute Post-Graduate Fellow and is supported by funding from NSW Ministry of Health, NSW Cancer Institute, Cancer Council of NSW, the Leukaemia Foundation of Australia and a research grant from MSD. This work was supported in part by the AIR@innoHK programme of the Innovation and Technology Commission of Hong Kong.

### Availability of data and materials

All data analyzed during this study are included in the published articles in Table 1. All analysis was done in R<sup>29</sup> version 4.0.3. The code to run treekoR is available on Bioconductor [40] and is available under the GPL-3 license. The version of source code used for the preparation of the manuscript and the benchmarking data generated in this study is available on Zenodo [41].

### Declarations

#### Ethics approval and consent to participate

Ethics approval is not applicable for this work.

#### Competing interests

The authors declare that they have no competing interests.

#### Author details

<sup>1</sup>School of Mathematics and Statistics, The University of Sydney, Sydney, New South Wales, Australia. <sup>2</sup>Charles Perkins Centre, The University of Sydney, Sydney, New South Wales, Australia. <sup>3</sup>Centre for Cancer Research, Westmead Institute for Medical Research, The University of Sydney, Sydney, New South Wales, Australia. <sup>4</sup>Faculty of Medicine and Health,

The University of Sydney, Sydney, New South Wales, Australia. <sup>5</sup>Blood Transplant and Cell Therapies Program, Department of Haematology, Westmead Hospital, Westmead, NSW, Australia. <sup>6</sup>Laboratory of Data Discovery for Health Limited (D24H), Science Park, Hong Kong SAR, China.

Received: 15 August 2021 Accepted: 26 October 2021

Published online: 29 November 2021

## References

1. Saeys Y, Van Gassen S, Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat Rev Immunol*. 2016;16(7):449–62. <https://doi.org/10.1038/nri.2016.56>.
2. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med*. 2018;50:96.
3. Marsh-Wakefield F, Ashhurst T. IgG B cells are associated with the development of multiple sclerosis. *Clin Transl Immunology*. 2020;9(5):e01133. <https://doi.org/10.1002/cti2.1133>.
4. De Biasi S, et al. Marked T cell activation, senescence, exhaustion and skewing towards TH17 in patients with COVID-19 pneumonia. *Nat Commun*. 2020;11(1):3434. <https://doi.org/10.1038/s41467-020-17292-4>.
5. Casneuf T, Adams HC III. Deep immune profiling of patients treated with lenalidomide and dexamethasone with or without daratumumab. *Leukemia*. 2021;35(2):573–84. <https://doi.org/10.1038/s41375-020-0855-4>.
6. Perfetto SP, Chattopadhyay PK, Roederer M. Seventeen-colour flow cytometry: unravelling the immune system. *Nat Rev Immunol*. 2004;4(8):648–55. <https://doi.org/10.1038/nri1416>.
7. Finak G, Langweiler M. Standardizing flow cytometry immunophenotyping analysis from the Human ImmunoPhenotyping Consortium. *Sci Rep*. 2016;6(1):20686. <https://doi.org/10.1038/srep20686>.
8. Newell EW, Cheng Y. Mass cytometry: blessed with the curse of dimensionality. *Nat Immunol*. 2016;17(8):890–5. <https://doi.org/10.1038/ni.3485>.
9. Qiu P, Simonds EF. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol*. 2011;29(10):886–91. <https://doi.org/10.1038/nbt.1991>.
10. Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. *Proc Natl Acad Sci U S A*. 2014;111(26):E2770–7. <https://doi.org/10.1073/pnas.1408792111>.
11. Van Gassen S, et al. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*. 2015;87(7):636–45. <https://doi.org/10.1002/cyto.a.22625>.
12. Levine JH, Simonds EF. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*. 2015;162(1):184–97. <https://doi.org/10.1016/j.cell.2015.05.047>.
13. Kiselev VY, Kirschner K. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods*. 2017;14(5):483–6. <https://doi.org/10.1038/nmeth.4236>.
14. Kim T, Chen IR. Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief Bioinform*. 2019;20(6):2316–26. <https://doi.org/10.1093/bib/bby076>.
15. Huang R, et al. treelimbR pinpoints the data-dependent resolution of hierarchical hypotheses. *Genome Biol*. 2021;22:1–21. <https://doi.org/10.1101/2020.06.08.140608>.
16. Tomic A, Tomic I, Dekker CL, Maecker HT, Davis MM. The FluPRINT dataset, a multidimensional analysis of the influenza vaccine imprint on the immune system. *Sci Data*. 2019;6(1):214. <https://doi.org/10.1038/s41597-019-0213-4>.
17. van der Laan MJ, van der Laan MJ, Pollard KS. A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *Journal of Statistical Planning and Inference*. 2003;117(2):275–303. [https://doi.org/10.1016/S0378-3758\(02\)00388-9](https://doi.org/10.1016/S0378-3758(02)00388-9).
18. Shen-Orr SS, et al. Defective signaling in the JAK-STAT pathway tracks with chronic inflammation and cardiovascular risk in aging humans. *Cell Syst*. 2016;3:374–384.e4.
19. Aghaepour N, et al. Critical assessment of automated flow cytometry data analysis techniques. *Nat Methods*. 2013;10(3):228–38. <https://doi.org/10.1038/nmeth.2365>.
20. Hu Z, Glicksberg BS, Butte AJ. Robust prediction of clinical outcomes using cytometry data. *Bioinformatics*. 2019;35(7):1197–203. <https://doi.org/10.1093/bioinformatics/bty768>.
21. Subrahmanyam PB, Dong Z. Distinct predictive biomarker candidates for response to anti-CTLA-4 and anti-PD-1 immunotherapy in melanoma patients. *J Immunother Cancer*. 2018;6(1):18. <https://doi.org/10.1186/s40425-018-0328-8>.
22. Teh CE, Gong JN. Deep profiling of apoptotic pathways with mass cytometry identifies a synergistic drug combination for killing myeloma cells. *Cell Death Differ*. 2020;27(7):2217–33. <https://doi.org/10.1038/s41418-020-0498-z>.
23. Weber LM, Nowicka M, Soneson C, Robinson MD. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Commun Biol*. 2019;2:183.
24. Weber LM, Robinson MD. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry A*. 2016;89(12):1084–96. <https://doi.org/10.1002/cyto.a.23030>.
25. Liu X, Song W. A comparison framework and guideline of clustering methods for mass cytometry data. *Genome Biol*. 2019;20(1):297. <https://doi.org/10.1186/s13059-019-1917-7>.
26. Krzak M, Raykov Y, Boukouvalas A, Cuttillo L, Angelini C. Benchmark and parameter sensitivity analysis of single-cell RNA sequencing clustering methods. *Front Genet*. 2019;10:1253. <https://doi.org/10.3389/fgene.2019.01253>.
27. Duò A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res*. 2018;7:1141.
28. Crowell HL, Zanotelli VRT, Chevrier S, Robinson MD. CATALYST: Cytometry dATa anALYSIS Tools; 2020.
29. R Core Team. R: a language and environment for statistical computing. (2020).
30. Bhattacharya S, Dunn P. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci Data*. 2018;5(1):180015. <https://doi.org/10.1038/sdata.2018.15>.
31. Krieg C, Nowicka M. High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. *Nat Med*. 2018;24(2):144–53. <https://doi.org/10.1038/nm.4466>.
32. Wagner J, et al. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell*. 2019;177:1330–1345.e18.

33. Geanon D, Lee B. A streamlined whole blood CyTOF workflow defines a circulating immune cell signature of COVID-19. *Cytometry A*. 2021;99(5):446–61. <https://doi.org/10.1002/cyto.a.24317>.
34. Neumann J, Prezzemolo T. Increased IL-10-producing regulatory T cells are characteristic of severe cases of COVID-19. *Clin Transl Immunology*. 2020;9(11):e1204. <https://doi.org/10.1002/cti2.1204>.
35. Mathew D, Giles JR. Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. *Science*. 2020;369(6508) <https://doi.org/10.1126/science.abc8511>.
36. Sade-Feldman M, Yizhak K. Defining T cell states associated with response to checkpoint immunotherapy in melanoma. *Cell*. 2019;176(1-2):404. <https://doi.org/10.1016/j.cell.2018.12.034>.
37. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139–40. <https://doi.org/10.1093/bioinformatics/btp616>.
38. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models Using lme4. *J Stat Softw*. 2015;67(1) <https://doi.org/10.18637/jss.v067.i01>.
39. Lang M, et al. mlr3: a modern object-oriented machine learning framework in R. *J. Open Source Softw*. 2019;4:1903.
40. Chan A, Patrick E. treekoR. Bioconductor; 2021. <https://doi.org/10.18129/B9.bioc.treekoR>.
41. Chan A, Patrick E. Adam2o1o/treekoR\_analysis: treekoR Manuscript Analysis. Zenodo. 2021; <https://doi.org/10.5281/zenodo.5591142>.
42. Bendall SC, Simonds EF, Qiu P, Amir EAD, Krutzik PO, Finck R, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*. 2011;332:687–96. <https://doi.org/10.1126/science.1198704>.
43. Gower JC, Ross GJS. Minimum spanning trees and single linkage cluster analysis. *Appl Stat*. 1969;18(1):54. <https://doi.org/10.2307/2346439>.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

