Identifying cellular-to-phenotype associations by elucidating hierarchical relationships in high-dimensional cytometry data

Authors

Adam Chan[1,2], Wei Jiang[3,4], Emily Blyth[3-5], Jean Yang[1,2], Ellis Patrick[1,3]

Affiliations

[1] School of Mathematics and Statistics, The University of Sydney, Sydney, New South Wales, Australia

[2] Charles Perkins Centre, The University of Sydney, Sydney, New South Wales, Australia

[3] Centre for Cancer Research, Westmead Institute for Medical Research, The University of Sydney, Sydney, New South Wales, Australia

[4] Faculty of Medicine and Health, The University of Sydney, Sydney, New South Wales, Australia

[5] Department of Haematology, Westmead Hospital, NSW Australia

*Corresponding author. Email: ellis.patrick@sydney.edu.au

Abstract

High-throughput single cell technologies hold the promise of discovering novel cellular relationships with disease and necessitate the use of effective analytical workflows. When manual gating is used to define cell types, the gating hierarchy can be used to identify cell types whose abundances change relative to a parent population. This strategy allows subtle changes to be observed that could be missed if small subsets were compared to all measured cells. However, typical analyses that employ unsupervised clustering overlook the valuable hierarchical structure present in cell type definitions by exclusively quantifying the proportions of cell type clusters relative to all cells. We present treekoR, a framework that facilitates 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 characterise cells by measuring up to fifty extracellular antigens[1], with single-cell sequencing platforms able to measure thousands of intracellular RNA[2]. Unfortunately, this ground-breaking capacity to deeply phenotype cells 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 nature of single-cell data, there exists an opportunity to exploit this structure 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 of a cell type relative to all cells in a sample (*%total*), or, as the proportion of a cell type relative to a parent population of cells (*%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 with either quad or polygon gates to identify cell populations in a sequential manner. For example, 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 makes it very natural to quantify the proportion of cell types relative to their parent lymphocyte population making the quantifications robust to drastic changes in unrelated subsets. The main drawbacks of this method however are its reliance on the time-consuming process of manually gating markers which has become impractical for high-parameter assays[8] and the substantial reliance on expert knowledge which may also tend to 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 utilised frequently in the analysis of high-dimensional cytometry data to identify cell populations. Whilst they have significantly improved the efficiency in which scientists can analyse 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, utilise minimal spanning trees over clustering as a visualisation tool. Citrus employs hierarchical clustering and regularised 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 whilst 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 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 main steps (**Figure 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) visualise 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 then helps to uncover important associations with a clinical end point of interest by visualising 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 (**Figure 2a**); we observed a reduction in CD4+ Tem cells in CMV positive patients using %parent (p=6.1×10^{-5}, FDR=3.33×10^{-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=1.5×10^{-3}, FDR=3.5×10^{-2}), but not with %total (p=0.26, FDR=0.69) (**Figure 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.

We applied treekoR to a flow cytometry panel of CD8+ T Cells, generated to distinguish between COVID-19 patients and healthy controls[4], and observed 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 (**Figure 2c**), we discovered a HLA-DR+ CD38+ CD8+ T cell whose %parent is greater in COVID-19 patients than healthy controls ($p=3.19\times 10^{-10}$, $FDR=2.76\times 10^{-8}$) (**Figure 2d**). 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 (**Figure 2e**), which is robust to the outliers observed in the %total quantification (**Figure 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

In several datasets, a greater discrimination is observed between the binary outcomes through quantifying proportions as %parent than %total. 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 clustering algorithms, HOPACH[17] and average-linkage hierarchical clustering, with both generating different estimates of %parent (**Supplementary Figure 1**). 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 (**Figure 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]. To determine if the incorporation of %parent features in multivariate classification models could help improve patient classification in several datasets, 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). There were various differences in balanced accuracy between using %total and %parent (using either HOPACH or hierarchical clustering with average linkage) in each dataset (**Figure 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 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 (**Supplementary Figure 1**). When each feature set was ranked using the mean and standard deviation of the balanced accuracy in each dataset (**Figure 4**), no single quantification of proportion performed the best for prediction of patient outcomes across all analysed cytometry datasets. The differences in rank however means 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 analysing 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 whilst 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. Whilst 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 analysing 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 amount of proportions calculated and potentially provide more biologically relevant %parent.

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 emphasise 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 main 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) visualise the significance results on the tree. Here we describe the steps in more detail, along with the parameters we used in the analyses throughout this paper.

(i) Clustering. Unsupervised clustering was performed using the FlowSOM[11] algorithm as part of the CATALYST[28] package in R[29], using a 10x10 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 analysed 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. 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. HOPACH allows for multiple children per node whilst other included methods only cater for two children per node. Other hierarchical clustering techniques can be used in treekoR and are included in the R stats hclust function[29]. Possible techniques include average-linkage, Ward-linkage, single-linkage, complete-linkage and McQuitty agglomerative hierarchical clustering.

(iii) Calculation of proportions. The proportions of the different cell clusters are then quantified once clustering and a hierarchical clustering tree of the clusters have been established in the data. 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, for each patient the proportions of cells belonging to the clusters in each node of the tree are measured as a proportion of the number of cells belonging to 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) Visualisation. The results of these proportions can be then visualised through a coloured 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 coloured on each node by the test statistic obtained when testing using %total of that node, with the branch connecting the child to the parent coloured 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 (COVID-19 T cells counted as two datasets - CD4 and CD8 T cells) and one single-cell RNA-seq datasets as shown in Table 1.

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-reponse to treatment	Subrahmanyam et al. 2018[21]
Anti-PD-1	CyTOF	Predicting response vs non-response in Anti-PD-1	85718	20	Response / Non-response to treatment	Kreig et al. 2018[31]

		treatment				
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-1	11591741 (60% of cells were sampled and analysed)	168	COVID-19 recovered / healthy	Mathew et al. 2020[35]

		9 vs. healthy				
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]

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.

Data normalisation

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, we performed p-value adjustment using the FDR method, whilst p-value adjustment was not performed in the benchmark comparison.

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[37]) 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.

All analysis was done in R[29] version 4.0.3.

Code availability

The code to run treekoR is available on Bioconductor (https://bioconductor.org/packages/release/bioc/html/treekoR.html) and code to reproduce the manuscript analysis from processed data has been shared on Github (https://github.com/adam2o1o/treekoR_analysis)

Author 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.

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. 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.

References

1. Saeys, Y., Van Gassen, S. & Lambrecht, B. N. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat. Rev. Immunol.* **16**, 449–462 (2016).
2. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 96 (2018).
3. Marsh-Wakefield, F. *et al.* IgG B cells are associated with the development of multiple sclerosis. *Clin Transl Immunology* **9**, e01133 (2020).
4. De Biasi, S. *et al.* Marked T cell activation, senescence, exhaustion and skewing towards TH17 in patients with COVID-19 pneumonia. *Nat. Commun.* **11**, 3434 (2020).
5. Casneuf, T. *et al.* Deep immune profiling of patients treated with lenalidomide and dexamethasone with or without daratumumab. *Leukemia* **35**, 573–584 (2021).
6. Perfetto, S. P., Chattopadhyay, P. K. & Roederer, M. Seventeen-colour flow cytometry: unravelling the immune system. *Nat. Rev. Immunol.* **4**, 648–655 (2004).
7. Finak, G. *et al.* Standardizing Flow Cytometry Immunophenotyping Analysis from the Human ImmunoPhenotyping Consortium. *Sci. Rep.* **6**, 20686 (2016).
8. Newell, E. W. & Cheng, Y. Mass cytometry: blessed with the curse of dimensionality. *Nat. Immunol.* **17**, 890–895 (2016).
9. Qiu, P. *et al.* Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature Biotechnology* vol. 29 886–891 (2011).
10. Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J. & Nolan, G. P. Automated identification of stratifying signatures in cellular subpopulations. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E2770–7 (2014).
11. Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and

interpretation of cytometry data. *Cytometry Part A* vol. 87 636–645 (2015).

12. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).
13. Kiselev, V. Y. *et al.* SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).
14. Kim, T. *et al.* Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief. Bioinform.* **20**, 2316–2326 (2019).
15. Huang, R. *et al.* treeclimbR pinpoints the data-dependent resolution of hierarchical hypotheses. doi:10.1101/2020.06.08.140608.
16. Tomic, A., Tomic, I., Dekker, C. L., Maecker, H. T. & Davis, M. M. The FluPRINT dataset, a multidimensional analysis of the influenza vaccine imprint on the immune system. *Sci Data* **6**, 214 (2019).
17. Laan, M. J. van der, van der Laan, M. J. & Pollard, K. S. A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *Journal of Statistical Planning and Inference* vol. 117 275–303 (2003).
18. Shen-Orr, S. S. *et al.* Defective Signaling in the JAK-STAT Pathway Tracks with Chronic Inflammation and Cardiovascular Risk in Aging Humans. *Cell Syst* **3**, 374–384.e4 (2016).
19. Aghaeepour, N. *et al.* Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods* **10**, 228–238 (2013).
20. Hu, Z., Glicksberg, B. S. & Butte, A. J. Robust prediction of clinical outcomes using cytometry data. *Bioinformatics* **35**, 1197–1203 (2019).
21. Subrahmanyam, P. B. *et al.* Distinct predictive biomarker candidates for response to anti-CTLA-4 and anti-PD-1 immunotherapy in melanoma patients. *J Immunother Cancer* **6**, 18 (2018).

22. Teh, C. E. *et al.* Deep profiling of apoptotic pathways with mass cytometry identifies a synergistic drug combination for killing myeloma cells. *Cell Death Differ.* **27**, 2217–2233 (2020).

23. Weber, L. M., Nowicka, M., Soneson, C. & Robinson, M. D. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Commun Biol* **2**, 183 (2019).

24. Weber, L. M. & Robinson, M. D. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry A* **89**, 1084–1096 (2016).

25. Liu, X. *et al.* A comparison framework and guideline of clustering methods for mass cytometry data. *Genome Biol.* **20**, 297 (2019).

26. Krzak, M., Raykov, Y., Boukouvalas, A., Cutillo, L. & Angelini, C. Benchmark and Parameter Sensitivity Analysis of Single-Cell RNA Sequencing Clustering Methods. *Front. Genet.* **10**, 1253 (2019).

27. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res.* **7**, 1141 (2018).

28. Crowell, H. L., Zanotelli, V. R. T., Chevrier, S. & Robinson, M. D. CATALYST: Cytometry dATa anALYSis Tools. (2020).

29. R Core Team. R: A Language and Environment for Statistical Computing. (2020).

30. Bhattacharya, S. *et al.* ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci Data* **5**, 180015 (2018).

31. Krieg, C. *et al.* High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. *Nature medicine* vol. 24 144–153 (2018).

32. Wagner, J. *et al.* A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. *Cell* **177**, 1330–1345.e18 (2019).

33. Geanon, D. *et al.* A streamlined whole blood CyTOF workflow defines a circulating immune cell signature of COVID-19. *Cytometry A* **99**, 446–461 (2021).

34. Neumann, J. *et al.* Increased IL-10-producing regulatory T cells are characteristic of severe cases of COVID-19. *Clin Transl Immunology* **9**, e1204 (2020).

35. Mathew, D. *et al.* Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. *Science* **369**, (2020).

36. Sade-Feldman, M. *et al.* Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell* **176**, 404 (2019).

37. Lang, M. *et al.* mlr3: A modern object-oriented machine learning framework in R. *J. Open Source Softw.* **4**, 1903 (2019).

38. Bendall, S. C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).

39. Gower, J. C. & Ross, G. J. S. Minimum Spanning Trees and Single Linkage Cluster Analysis. *Applied Statistics* vol. 18 54 (1969).

Figures and Tables

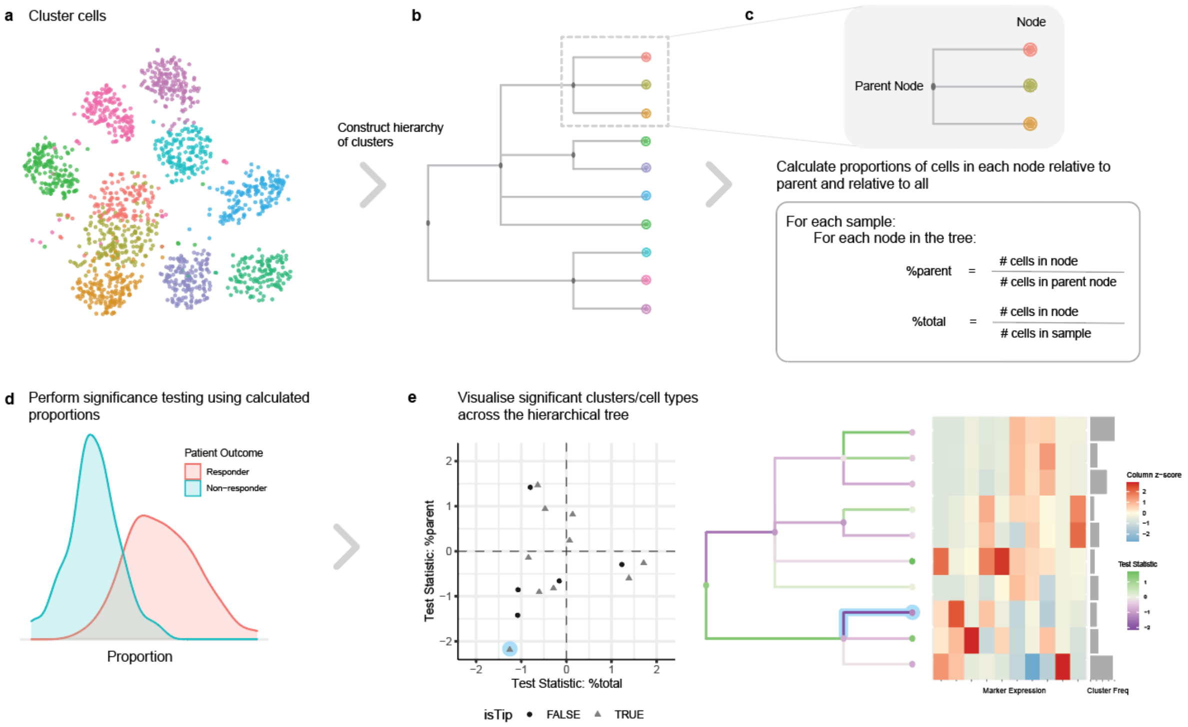

a Cluster cells
b
Construct hierarchy of clusters
c
Node
Parent Node
Calculate proportions of cells in each node relative to parent and relative to all
For each sample:
For each node in the tree:
%parent = # cells in node / # cells in parent node
%total = # cells in node / # cells in sample
d Perform significance testing using calculated proportions
Patient Outcome
Responder
Non−responder
Proportion
e Visualise significant clusters/cell types across the hierarchical tree
Test Statistic: %parent
2
1
0
−1
−2
−2
−1
0
1
2
Test Statistic: %total
isTip ● FALSE ▲ TRUE
Column z−score
2
1
0
−1
−2
Test Statistic
1
0
−1
−2
Marker Expression
Cluster Freq

Fig 1: treekoR helps to extract insight from cytometry data through deriving a hierarchy of cell clusters and measuring proportions to parent. **a.** An example t-SNE plot showing clustering of single cell data.**b.** Hierarchical tree constructed using HOPACH algorithm on the cluster median marker expressions. **c.** Definition of proportions to parent and proportions to all defined according to the organisation of the hierarchical tree. **d.** Significance testing is performed using both types of proportions calculated, testing for difference between the patient clinical endpoint of interest. **e.** Visualisation of the significance testing results. On the left, a scatterplot of each node in the hierarchical tree with the test statistic calculated using the %total (x-axis) vs. the test statistic calculated using the %parent (y-axis). On the right of the scatterplot, the hierarchical tree is coloured with the test statistics: the nodes coloured by the test statistic using %total and the branches of the nodes coloured by the test statistic using %parent. An example of a corresponding node between the two graphs is highlighted in blue. The heatmap plots the median marker expression of the leaf nodes to assist in identification of the corresponding cell clusters.

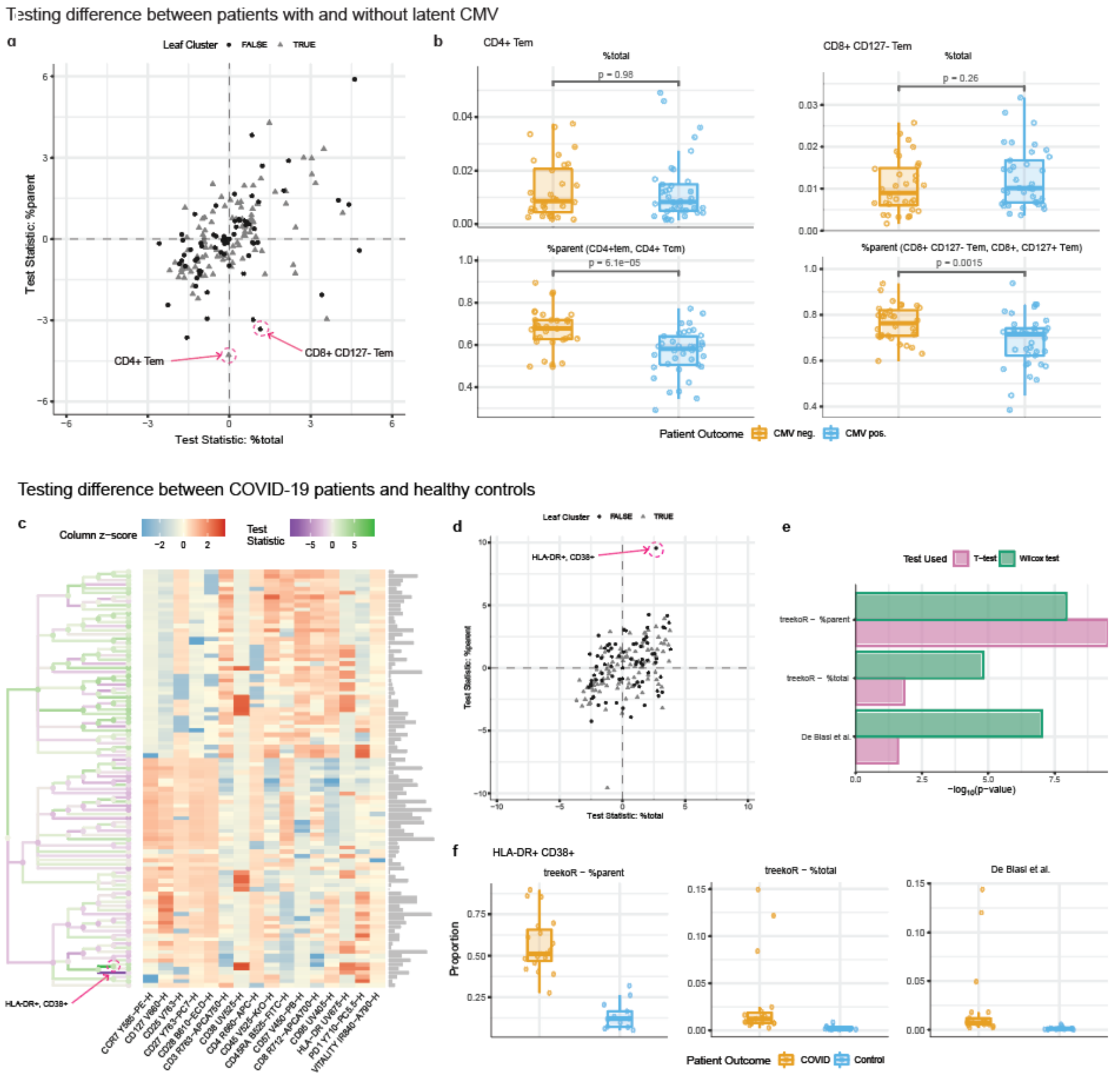


Fig 2: Measuring %parent can provide additional insight over %total. **a.** Scatterplot of test statistics with the cell clusters in differentiating between latent CMV infection patients. Highlighted clusters are significant using %parent, whilst not significant using %total. **b.** Comparative boxplot 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 coloured by the test statistic using the corresponding %total (nodes) and %parent (branches). The heatmap is coloured by the scaled cluster median expression values characterise 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 -log10 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

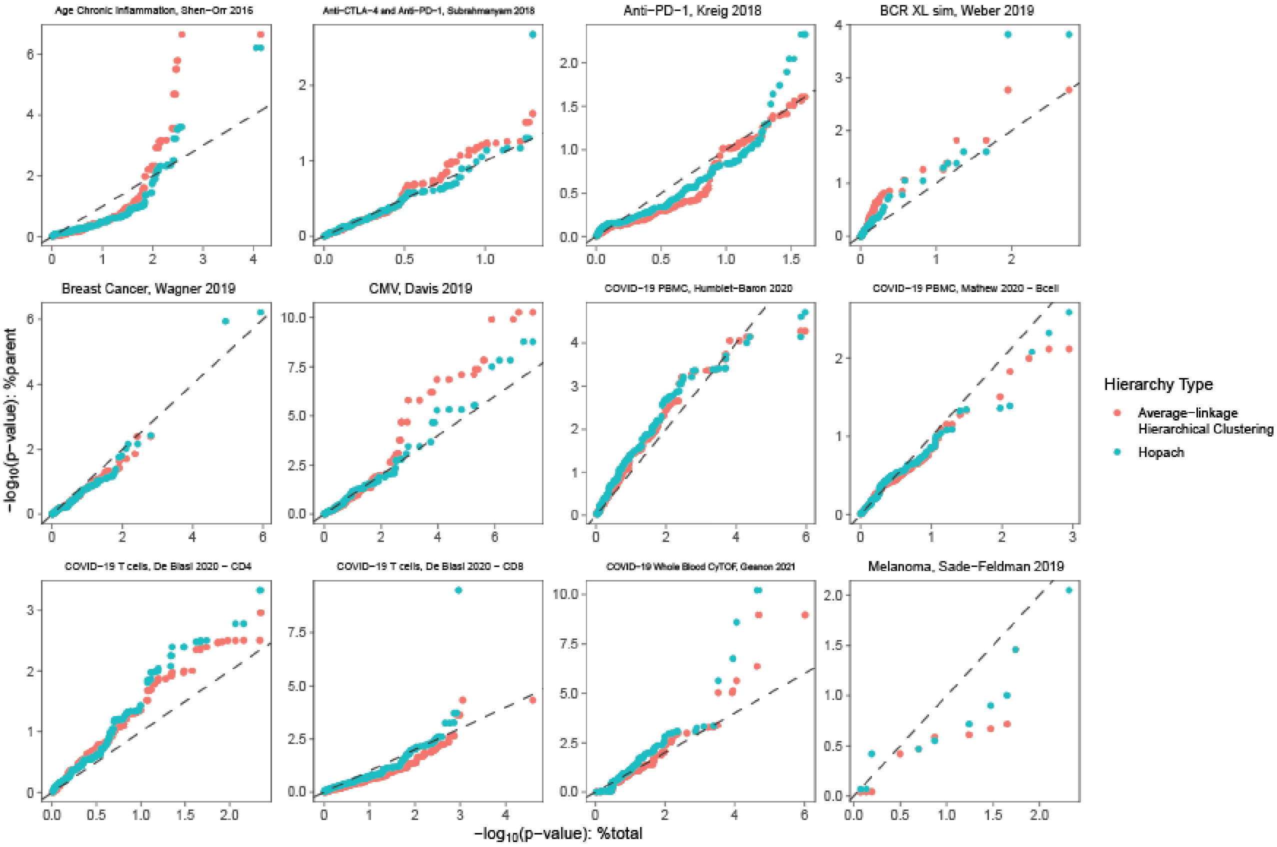


Fig 3: treekoR provides stronger associations with patient clinical outcomes. Cell clusters, constructed using both average-linkage hierarchical clustering and HOPACH, were tested between patient conditions using %total and %parent. Q-Q plots were plotted for each dataset by plotting the ordered negative log p-values using %total (x-axis) vs. using %parent (y-axis).

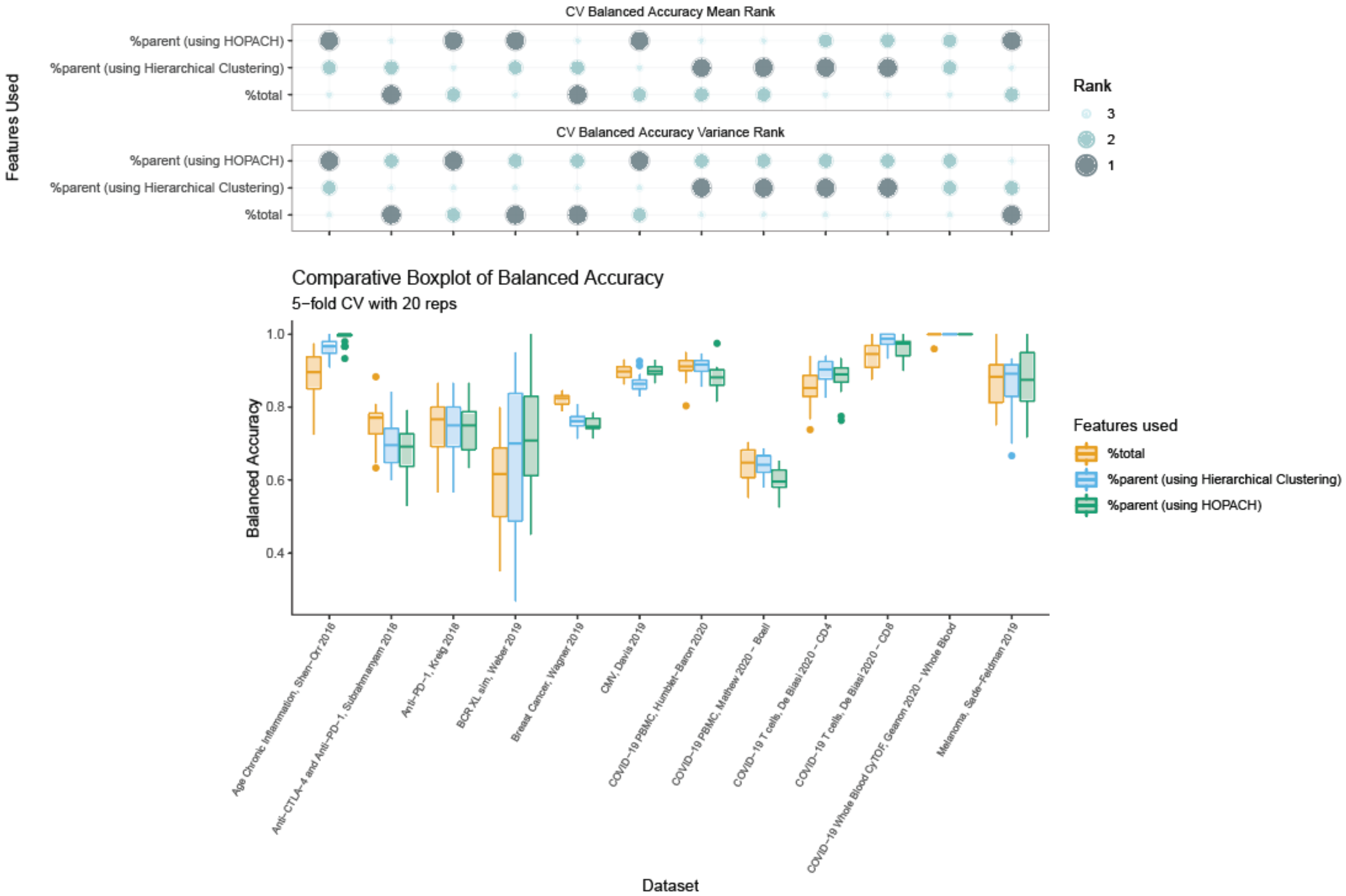


Fig 4: Measuring %parent offers improvements in patient classification performance. **a.** Comparative boxplots (lower panel) of balanced accuracy rates for each dataset and feature set: %total, %parent using average-linkage hierarchical clustering, and %parent using HOPACH. Values plotted are from a 5-fold CV with 20 repetitions, averaged across each repetition. The rank of each feature set within each dataset is shown in the bubble plot (upper panel), with rank 1 being the best (highest mean / lowest variance) and rank 3 being the worst (lowest mean / highest variance).

Supplementary

Different hierarchical clustering methods uncover varying %parent relationships in cytometry data

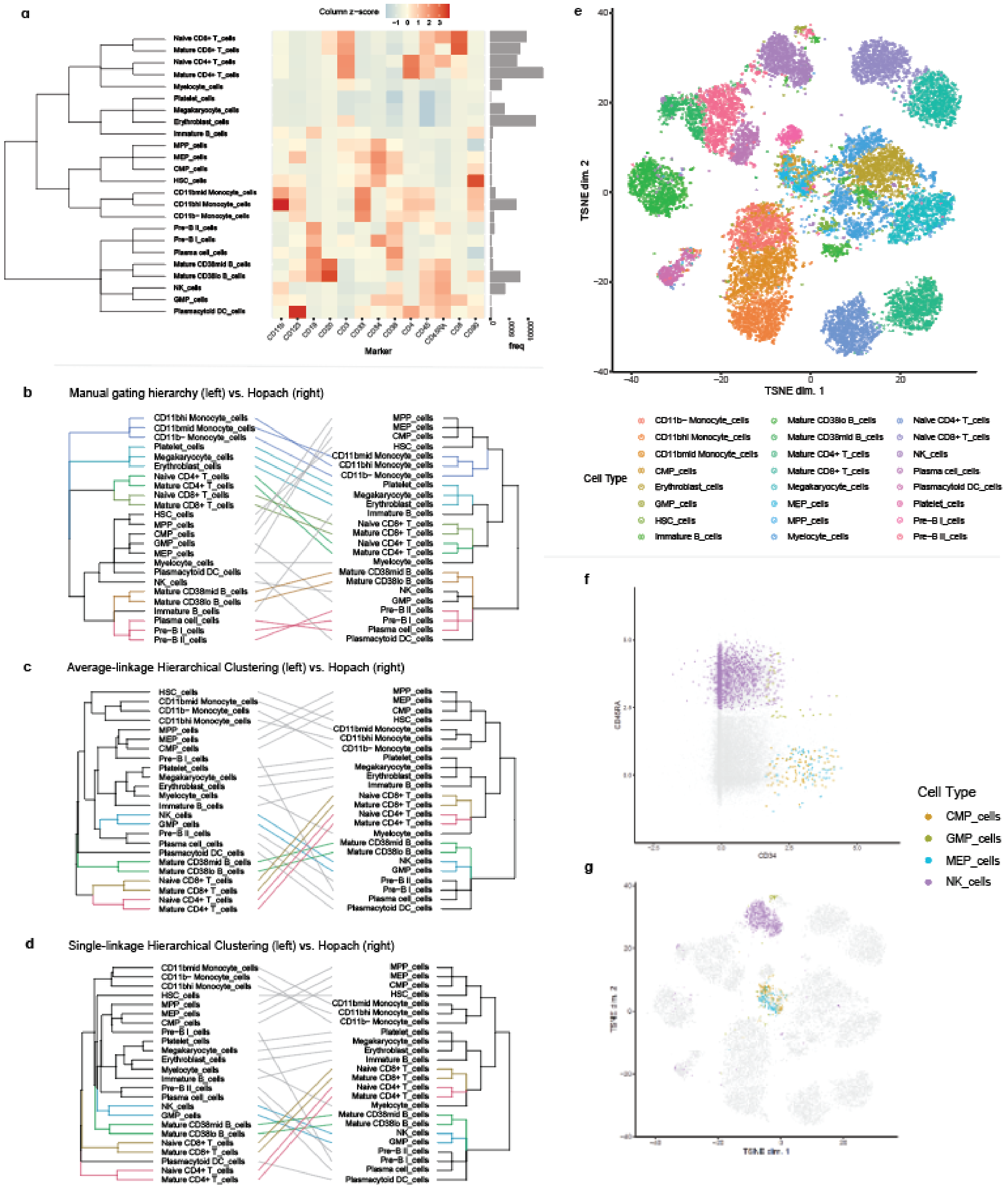
In this section we compare (in more detail) several tree structures derived from different hierarchical aggregation techniques, in addition to a manual gating tree structure, and how these ultimately result in capturing different signals from the data. Although no single representation may be the absolute correct one, exploring these different representations can begin to help analysts to uncover a broader scope of complex relationships that exist within cytometry data to discover the cellular heterogeneity between patient samples.

In our framework, we use the HOPACH algorithm to hierarchically aggregate the clusters into a tree. HOPACH is a clustering algorithm that was originally developed for gene expression data analysis, but has some properties which make it useful for the analysis of high-dimensional cytometry data. One advantage is the lack of restriction for splits to be binary, allowing up to 15 child nodes per node (**Supplementary Figure 1a**). In this example it can be observed that Pre-B cells, Mature B Cells and Immature B Cells fall under the same parent node, which, for example, would allow analysts to determine whether the compositional makeup of B cells plays any role in patient disease. Another advantage of HOPACH is the cluster collapsing step, which helps to alleviate any incorrect splitting of clusters. This helps prevent the tree from containing too many branches, which can reduce some correlations between proportions as our framework explores each of the parent-child relationships in the generated trees.

The high-dimensional nature of single cell cytometry data gives rise to numerous biologically relevant cell type hierarchies. treekoR acknowledges this by providing a framework which is not restricted to one specific cell type hierarchy constructed by a specific algorithm. We compared trees constructed using Hierarchical Ordered Partitioning And Collapsing Hybrid (HOPACH)[17] clustering, average-linkage hierarchical clustering, and single-linkage hierarchical clustering via ‘tanglegrams’ (a pair of trees drawn with edges connecting matching leaves between the pair) on a PBMC sample from a healthy bone marrow donor[38] (**Supplementary Figure 1b-1d**). The comparison highlights distinct cell type trees, which would consequently result in distinct quantifications of %parent proportions. Average-linkage hierarchical clustering - used in algorithms such as treeclimbr and citrus - and single-linkage clustering - closely resembling minimum spanning trees[39] used in SPADE and visualising FlowSOM - generate distinct hierarchical representations of the data. In this dataset, HOPACH provided a representation more closely resembling the manual gating tree constructed by Bendall et al. Of importance is that one representation of cell type hierarchy may not necessarily be the most informative, but each of these representations can lead to diverse yet relevant %parent relationships.

When comparing the HOPACH constructed tree to the manual gating hierarchy, a clear difference is how NK cells and GMP cells are grouped - NK and GMP fall under the same immediate parent node in the HOPACH tree whilst they are not in the gating tree (**Supplementary Figure 1b**). An analyst may gate the cell types differently using two markers such as CD34 & CD45RA (**Supplementary Figure 1f**). Although this provides a more interpretable representation to cell groupings, it is clear in the t-SNE plot

(**Supplementary Figure 1g**) how they would group together via clustering which considers all the available markers in unison. This indicates the effect of the experimental panel design, where some biologically distinct cell subpopulations can group together in automated clustering methods when there are more markers to distinguish them. Although there are notable differences between the automatically generated tree and the manually gated hierarchy, the HOPACH clustering is able to regenerate some of the cell type groupings. Despite some scenarios where the parent proportions may make less sense, this provides our framework with a big advantage over manual gating through being more efficient in handling larger datasets as well requiring a lower extent of prior knowledge to subset cells.



b Manual gating hierarchy (left) vs. Hopach (right)

c Average-linkage Hierarchical Clustering (left) vs. Hopach (right)

d Single-linkage Hierarchical Clustering (left) vs. Hopach (right)

Supplementary Fig 1: treekoR facilities varying cell type hierarchies to enable the measurement of cell types %parent a. A heatmap of the median marker frequencies, clustered using HOPACH, for each cell type as subsetted with manual gating from a healthy human bone marrow sample[38] (top left). **b.**Tanglegrams comparing two trees to highlight differences in cell type hierarchies: manual gating hierarchy vs. HOPACH clustering; **c.** average-linkage hierarchical clustering vs. HOPACH clustering; and **d.** single-linkage hierarchical clustering vs. HOPACH clustering (bottom right). **e.** A t-SNE plot of the sample, highlighted by the manually gated cell types **f.** Scatterplot of CD34 vs. CD45RA with only CMP, GMP, MEP & NK cells highlighted **g.** A t-SNE plot of the sample with only CMP, GMP, MEP & NK cells highlighted