Identifying differential cell populations in flow cytometry data accounting for marker frequency

Alice Yue[1], Cedric Chauve[2,3⋆], Maxwell Libbrecht[1]*, and Ryan Brinkman[1,4,5]*

[1] School of Computing Science, Simon Fraser University
[2] Department of Mathematics, Simon Fraser University
[3] LaBRI, University of Bordeaux
[4] Terry Fox Laboratory, BC Cancer Agency
[5] Department of Medical Genetics, University of British Columbia

Abstract. Bio-markers are measurable indicators that predict a given phenotype or disease. We introduce a method capable of discovering biologically meaningful, interpretable, and actionable differential cell population bio-markers from flow cytometry samples of different phenotypes. Cell populations are groups of cells that contain the same set of proteins. Differential cell populations are those that have a significantly changed abundance between samples of different phenotypic types. Existing methods for differential cell population identification fall into one of three categories: methods that 1) compare a limited set of pre-specified mutually exclusive cell populations that do not share cells, 2) find differential cell populations as a byproduct of another procedure, and 3) compare overlapping cell populations in a search space of all possible cell populations. The cell populations analyzed in 3) are dependent on each other and can be difficult to interpret. For example, an increase in one cell population (e.g. a bio-marker for a phenotype of interest) may induce an increase in several cell populations that share its cells. Our method solves this issue by taking into account these dependencies by finding only cell populations that are the source of these changes. Bio-markers can then be interpreted via a lattice-based visualization tool that depicts how these bio-markers affect each cell population and how they differentiate between samples of different phenotypes.

Keywords: Flow cytometry · Statistical Analysis · Bio-marker Identification.

⋆ Corresponding author.

1 Introduction

Bio-markers are measurable indicators that predict a given phenotype or disease. In this work, we introduce a method capable of discovering biologically meaningful, interpretable, and actionable bio-markers from flow cytometry (FCM) samples of different phenotypes.

A *flow cytometer* is a high-throughput apparatus capable of simultaneously measuring more than 40 properties per single cell, for samples containing hundreds of thousands of cells [1,8,9]. Flow cytometers are routinely used to study immune diseases (e.g. leukaemia and lymphoma [25]). FCM experiments can be split into two parts: physical detection of cell characteristics via a *cytometer* and analysis of these characteristics for cell sorting (i.e. cell population identification). A biological sample is mixed with a cocktail of *markers*, each of which attaches onto a target protein on a cell and emits a *fluorescence* of a certain wavelength when excited by a laser in the flow cytometer. This fluorescence is detected by an array of photomultiplying tubes, each measuring the *fluorescence intensity (FI)* of a certain range of wavelength. A high FI detected of a certain marker is then indicative of whether or not a cell contains the corresponding type of protein. In an attempt to increase the amount of markers that can be simultaneously measured, newer machines, such as the mass cytometer (CyTOF) and mass synaptometers (SynTOF), measure non-FI properties, such as the mass-to-charge ratio of isotopes. Nevertheless, the output of these machines can be standardized and analyzed in a similar manner. An FCM experiment generates for each analysed sample an FCM Standard format [23] file containing an $R \times L$ matrix containing FI values. R is the number of cells while L is the number of markers. FCMs also detect a cells' size (forward scatter FS) and granularity (side scatter SS) but we will refer to the columns of this matrix as markers for simplicity.

Cell populations are groups of cells that contain the same set of proteins. Hence, each cell in an FCM matrix can be labelled as being a part of a cell population based on whether or not it shows high FI for a specific set of markers. Cell population identification in FCM is used in both clinical and research settings [20,17]. For example, FCM samples from patients can be classified into different **classes** ('control' or 'experiment(s)'; e.g. healthy or diseased) based on the values (e.g. higher than normal) of some descriptor 'feature type' (e.g. cell count) for some cell population(s). In this work, we aim to identify cell populations whose abundance differ significantly between samples of different phenotypes. We define these as *differential cell populations* or *bio-markers*.

While analysis of FCM samples can be and is predominantly done manually, computational alternatives for FCM analysis are being developed and used — motivated by the inefficiency, human error, subjectivity, and bias [22,12,17,15] originating from manual analysis [6]. In the present work we are interested in methods that classify FCM samples based on their phenotypic properties. To create such classifiers, one would use samples already labelled with a phenotype class to (1) train classification model(s) to become working classifiers (e.g. [11,10,4]); and (2) identify the main features or bio-markers that differentiate between these samples (e.g. [27,26,24,18,7,14]). The latter option provides results that can be used to classify future samples in an interpretable manner while the former outputs a black box. Hence, our method is one of the latter that focuses on finding differential cell poulations. Existing methods for differential cell population identification fall into one of three categories. The first category compares pre-specified cell populations. These methods are unable to properly classify samples whose phenotype is linked to unspecified cell populations. The second category finds differential cell populations as a byproduct of another procedure. For example, CytoDx [11] focuses on classifying FCM samples using regression directly over individual cells. As a means to interpret this regression model, it provides the regression weights for each cell and its FI values as input into a decision tree from which it obtains a set of important markers representative of differential cell populations. A shortcoming of these methods is that they explore only mutually exclusive cell populations (i.e. cell populations that do not share cells) and are based on an original

model not designed for differential cell population identification, making the resulting interpretation ad hoc. The third category, implemented by flowType [18], considers each possible subset of markers as defining a potential differential cell population. However, the proportion of cells in these cell populations are dependent on each other and can be difficult to interpret. For example, an increase in one cell population (e.g. a bio-marker for a phenotype of interest) may induce an increase in several cell populations that share a subset of its cells.

Our method solves these issues by taking into account these dependencies among proportions of related cell populations. More precisely, we introduce a measure of cell population abundance that controls for inter-cell population relationships. We show that our method is able to find robust, biologically meaningful, and interpretable bio-marker cell populations using visualization tools that depict how these bio-markers affect each cell population and how they differentiate between samples of different phenotypes. The rest of the paper is organized as follows. First, we will go into the FCM data processing pipeline and the methodology behind extracting our feature type. Thereon-after, we compare the results of the differential cell population extraction methods to evaluate the robustness and effectiveness of our feature type in bio-marker discovery on real and simulated data.

2 Method

2.1 Quantifying abundance of cell populations in an FCM experiment

Taking a pre-processed FCM sample matrix as input (see standard FCM sample pre-processing protocol in [19]), we first convert this into a vector of cell population counts representing abundance of cells. There are over 50 tools developed to identify cell populations [13] categorized as (1) supervised classification, (2) unsupervised clustering, and (3) automated gating (termed a parent type of unsupervised clustering) of cells. While (1) and (2) are applications of classification/clustering tools on cells as objects, automated gating mimics the manual process of gating, a computerized way of a traditionally manual process of drawing borders or gates around cell populations on 2D scatterplots, two markers at a time — as specified by an expert created *gating strategy*. In our method, we use (3) to obtain threshold gates using flowDensity [16] and to exploit expert knowledge in the form of a gating strategy. Gates specify the value from which an FI value can be termed "high" (i.e. a cell expresses the corresponding marker). This step outputs L marker threshold gates per sample. These gates and their corresponding pre-processed FCM samples ($R \times L$ matrix) are then given as input to flowType [18] which generates the cell count for each of the $m = 3^L$ possible cell populations, each being labelled based on whether or not a specific group of markers have high FI or are present on the cells. This allows us to explore the entire search space of all cell populations. Each cell population is labeled by a set of $0 \leq \ell \leq L$ unique markers that it expresses positively$^+$ or negatively$^-$ for (i.e. has FI above or below the corresponding gate. The cell counts for n samples are then collated into a $n \times m$ matrix.

2.2 Cell population hierarchy

To model and visualize the m cell populations in each FCM sample, we use the *cell hierarchy*. A cell hierarchy is a directed acyclic graph where nodes represent cell populations and arcs represent maximal inclusion relation between cell populations. We define the ℓ'th level of the hierarchy as the set of all nodes whose label contains exactly ℓ markers. Each arc specifies and links together a 'parent' and its maximal 'child' sub-population(s) defined by the addition of one 'marker condition' (e.g. A^+ and A^- are positive and negative marker conditions respectively). For example, if there are three markers $\{A, B, C\}$, then there are arcs from the node representing the cell population of label

A^+ to the nodes of labels A^+B^+, A^+C^+, A^+C^- and A^+B^-. Note that A^+ represent the cells that have a high FI for marker A irrespective of its FI for other markers. We refer to Figure 1 for an illustration.

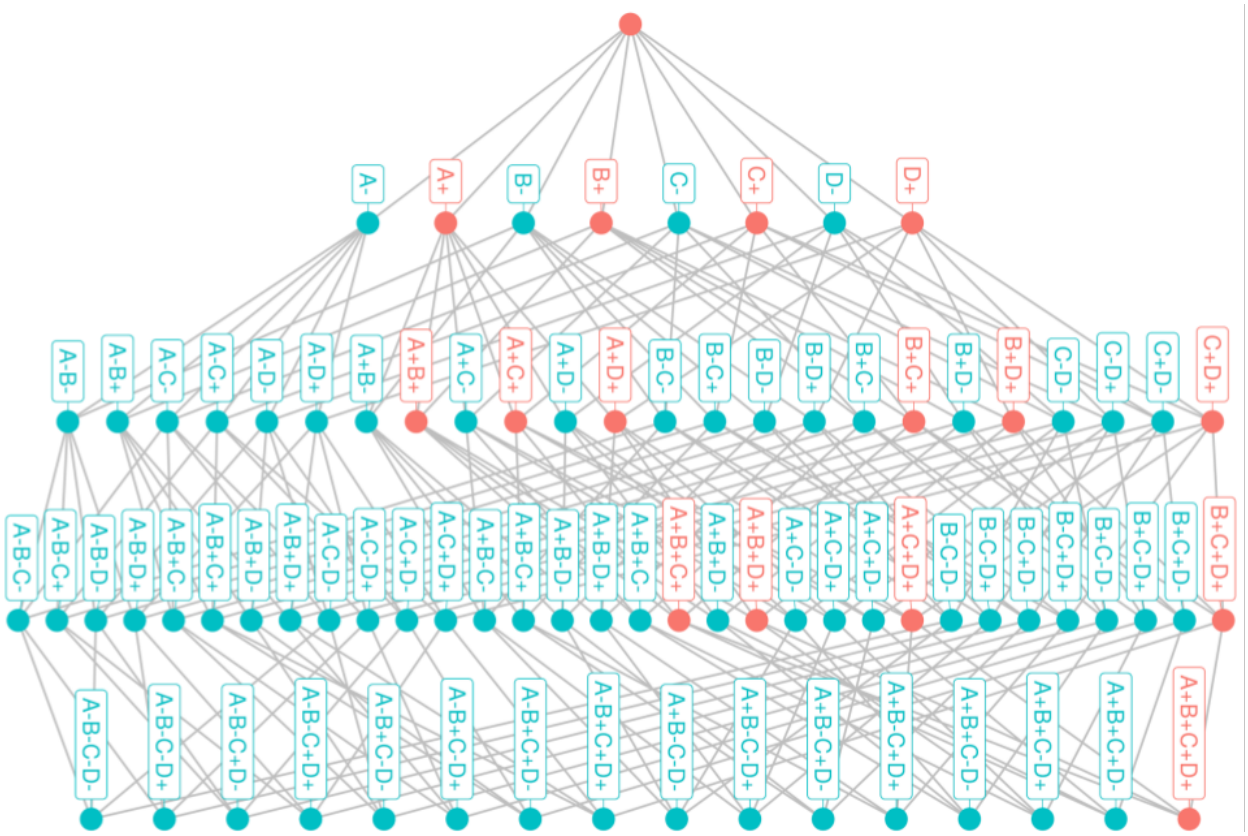

Fig. 1. An example of a cell hierarchy representation of an FCM sample and its cell populations defined by markers A, B, C, and D.

2.3 Features quantifying cell population abundance

Aiming to classify samples according to phenotypic differences, our key goal is to identify possible biomarkers or differential expressed cell population whose cell abundance properties differ significantly between samples representing different phenotype(s).

In this work, we compare several features that quantify cell population abundance. The simplest such feature is the raw count of cells in each population ('count'). These counts are typically converted into proportions ('prop') or the count of each node over the count of the root node representing all cells. Using proportions, however, introduces a situation where if only one cell population changes (e.g. A^+'s count increases to have proportion .67), then it will cause an artificial change in other related cell populations (e.g. A^-'s count remains unchanged but its proportion becomes .33). Therefore, another way of normalizing the cell count is to use the trimmed mean (TMM) [21] to get normalized counts ('countAdj'). Assuming that most cell population abundances do not change between samples of different phenotypic types, TMM adjusts the cell counts such that those unchanged cell populations are not artificially enriched.

2.4 A novel feature type: specific enrichment ('SpecEnr')

Expected proportion The novel feature type we introduce, rather than using proportions directly, revolves around the idea of detecting whether a change in a cell population proportion is the result of its actual change or an after-affect of changes in its descendant(s) cell populations (child, and child of child) or another cell population on layer ℓ — i.e. cell populations on layer $\geq \ell$. For this, we calculate the *expected proportion* of a cell population in layer ℓ given the actual proportions of its ancestor (parent, and parent of parents) cell populations.

We first denote the proportion P of a node in layer ℓ by $P(v^{1:\ell})$ (e.g. $P(v^{1:3}) = P(A^+B^+C^-)$) where $1:\ell$ represents the indices of the markers its label contains. First, we assume $P(v^1)$ (e.g. A^+) and $P(v^2)$ (e.g. B^+) are independent given $P(v^{3:\ell})$ (e.g. C^-).

$$P(v^1|v^{2:\ell}) = P(v^1|v^{3:\ell}) \tag{1}$$

$$P(v^{1:\ell}) = P(v^{2:\ell})\frac{P(v^1, v^{3:\ell})}{P(v^{3:\ell})} \tag{2}$$

Generalizing our assumption to any p, q pair, we get

$$P(v^{1:\ell}) = P(v^{1:\ell\setminus p})\frac{P(v^{1:\ell\setminus q})}{P(v^{1:\ell\setminus\{p,q\}})} \tag{3}$$

where $p \in 1:\ell$ and $q \in 1:\ell \setminus p$. We also note that as the above requires $P(v^{1:\ell\setminus\{p,q\}})$ to exist, we note the expected proportion is only calculated for cell populations in layers $\ell \geq 2$. We initialize the root and layer one nodes' expected proportions as 1 and .5 respectively.

In Equation 3, we assumed that all marker pairs q, p should be independent of each other, therefore any pair of p, q should yield a $P(v^{1:\ell})$ whose value equate to v's actual proportion. However, suppose A^+C^+'s cell proportion increased while B^+'s proportion remained the same and these non/changes happened independent of each other. While A^+ and C^+ are dependent on each other, B^+ is independent of both A^+ and C^+. In the latter case, the initial assumption we made with the scenario in Equation 2 still holds if we are calculating the expected proportion of $A^+B^+C^+$ — that is $P(A^+)$ and $P(B^+)$ are independent given $P(C^+)$. In this scenario, we also imply that the only differential expressed or changed cell population here is A^+C^+. As such, we formally define a *differential expressed* cell population $v^{1:\ell}$ where $P(v^p)$ is dependent on $P(v^q)$ for all $\{p, q\} \in 1:\ell$ (i.e. Equation 3 will not hold for any p, q). However, this also applies to direct ancestors of the differential expressed cell population since they are defined on a subset of markers defining the differential expressed population. Hence, a second part to this definition is that the differential expressed cell population should not have any direct descendants who are also differential expressed. Given our definition, we relax our assumption in Equation 3 by stating that if the cell population we are calculating an expected proportion for is not differential expressed or is an ancestor of a differential expressed cell population, there must be some p, q pair such that $P(v^p)$ is independent of $P(v^q)$. One way to obtain this p, q pair is via the following (see Appendix 2 for proof of correctness).

$$P(v^{1:\ell}) = \max_{p\in 1:\ell} P(v^{1:\ell\setminus p}) \min_{q\in 1:\ell\setminus p} \frac{P(v^{1:\ell\setminus q})}{P(v^{1:\ell\setminus\{p,q\}})} \tag{4}$$

Note that the min and the max in the equation above can be switched and the result will remain the same because p and q are mutually independent.

Specific enrichment ('SpecEnr') Assuming we have calculated the expected proportions for all cell populations (see paragraphs below for the algorithmic details), our specific enrichment feature type (SpecEnr) then questions whether the actual observed proportions differ from the expected proportions by taking the natural log of the actual over the expected proportions.

$$SpecEnr(v) = \ln \frac{P'(v)}{P(v)} \tag{5}$$

To extract significantly differential feature values between samples of different classes, we use the T-test to obtain our p-values for each cell population which we adjust to obtain q-values using the Benjamini–Yekutieli (BY) [5] procedure to control for the false discovery rate (unless specified otherwise).

Note that cell hierarchy plots we will show in the results section only colours cell populations that have a significant q-value below .05.

Algorithmic details To reduce runtime, we only directly calculate this expected proportion for cell populations with only positive marker conditions (marked in red on Figure 1). This runtime includes operations to first calculate all the arc values in our cell hierarchy. Arc values can be interpreted as the proportion of the cell population an arc points to over the proportion of the cell population an arc originates from. Since there are $\sum_{\ell=0}^{L} \left\{ l \cdot \binom{L}{\ell} \right\} = 2^{L-1}L$ edges, this step takes $O(2^L)$ operations. Next, to find the maximum parent proportion and minimum value on edges pointing to the non-maximum parent, we do ℓ and $\ell(\ell-1)$ operations for each cell population. Given that there are $\binom{L}{l}$ cell populations with only positive marker conditions on each layer, this part requires $\sum_{\ell=2}^{L} \ell^2 \binom{L}{l}$ or $O(\ell^3)$ operations. If we were to calculate the above for all cell populations, the first part would require $O(3^L)$ as opposed to $O(2^L)$ operations and we would have to consider $2^\ell \binom{L}{l}$ instead of $\binom{L}{l}$ cell populations per layer for the second part.

To follow up on all other cell populations, we also state that the feature values for these cell populations collectively implicitly contain information about all nodes in the full hierarchy. As such, we can deduce the expected proportion of all other cell populations with at least one negative marker condition. For example, the first layer of a cell hierarchy can be specified with L cell populations. The count of these cell populations with a negative marker condition (e.g. A^-) can be deduced by the difference between the total proportion and the proportion of the corresponding cell population expressing positively for that marker (e.g. $P(A^-) = P(root) - P(A^+)$). The L'th layer can be specified with one cell population i.e. a cell population with all markers. The total number of cell populations needed to specify all cell populations is 2^L or the number of cell populations with only positive marker conditions.

From this, we use Algorithm 1 to infer the expected proportion of all other cell populations with at least one negative marker condition. This algorithm takes one difference operation per cell population amounting to $O(3^L)$ operations total. Since we only calculate expected proportions for $O(2^L)$ cell populations and since the total number of other cell populations far out-scales $O(2^L)$, we achieve an overall runtime of $O(3^L)$.

In the case that there is more than one threshold per marker (e.g. on top of A^- and A^-, there also exists $A^+ +$, $A^+ + +$, and so on), the proof of correctness for expected proportions still holds as

a cell population in layer ℓ must have ℓ parent nodes. Algorithm 1 also still holds; the only difference is that the proportions represented by index $p+$ is now not a single proportion value but a sum of all proportions with a positive marker condition for the marker on index p. Note that this is contingent on a characteristic of flowType who labels markers with additional thresholds as positive marker conditions. Conversely, there can be only one negative marker condition per markers (i.e. there can be no A^-−). As we still only calculate expected proportion for cell populations containing exclusively positive marker conditions, the runtime of expected proportion calculation increases with the number of thresholds but the overall runtime remains the same. If each marker contains c thresholds, then the number of cell population containing only positive marker conditions increase from $O(2^L)$ to $O(c \cdot 2^L)$ — equivalently $O(2^L)$.

Data: $V \leftarrow$ cell populations with all positive marker conditions whose expected proportion is already calculated.
Data: $V^- \leftarrow$ all cell populations with ≥ 1 negative marker condition(s).
for $\ell := 2 \rightarrow L$ **do**
  $\mathbf{v}_\ell \leftarrow \{v | v \in V, |v| = \ell\}$ **while** $v_\ell \neq \{\}$ **do**
    $\mathbf{v}^*_\ell \leftarrow \{\}$
    for $v_\ell := \mathbf{v}_\ell$ **do**
      for $p^+ :=$ a marker in v_ℓ with a positive$^+$ condition **do**
        $P(v^{p^-,1:\ell\setminus p^+}) = P(v^{1:\ell\setminus p^+}) - P'(v_\ell)$
        $\mathbf{v}^*_\ell \leftarrow \{\mathbf{v}^*_\ell, v^{p^-,1:\ell\setminus p^+}\}$
        $V = \{V, v^{p^-,1:\ell\setminus p^+}\}$
        $V^- = \{V^- \setminus v^{p^-,1:\ell\setminus p^+}\}$
      end
    end
    $\mathbf{v}_\ell = \mathbf{v}^*_\ell$
  end
end

Algorithm 1: Calculating the expected proportion of cell populations with negative marker conditions

2.5 Experiment data

Synthetic data To test the robustness of our method, we synthetically generate a negative control data set (ctrl) with no bio-markers/difference between samples. It contains two classes with 500 samples each containing 300,000 cells analyzed on 4 markers (A, B, C, and D). For each cell, we randomly assign it to be positive or negative for each marker with 0.5 probability. Therefore, each cell population at hierarchy depth ℓ has proportion around $.5^\ell$.

To visualize whether or not our method can capture known bio-markers, we also generate synthetic positive controls in the same way as the negative controls except (1) in data set 'pos1', $A^+ = .25$, $A^- = .75$. To see how our method behaves when there are multiple changed cell populations, we also generate data set (2) 'pos2', where $A^+ = .25$, $A^- = .75$ and $B^+ = .25$, $B^- = .75$. We also test if our method works when there are multiple changed cell populations at lower levels of the cell hierarchy by generating (3) 'pos3' whose A^+B^+ increases by 50% and (4) 'pos4' whose A^+B^+ and C^+D^+ increases by 50%. We also show and discuss how a user can interpret and analyze the cell hierarchy when cell populations who are ancestor/descendants of each other changes. For this, we have data set (5) 'pos5' whose A^+ and $A^+B^+C^+$ both increase by 50%. Note that the results we obtain is equivalent when the same scenarios are given but the cell population decrease in proportion (Figure 5 in the Appendix).

Real data We also tested our method on real data sets to find existing and discover new bio-markers.

We use a data set from the FlowCAP-II challenge [2] containing two classes: samples from 316 healthy and 43 AML (acute myeloid leukemia) positive subjects' blood or bone marrow tissues — each with around 60,000 cells. We use the 6th 'panel' or set of markers ($HLA-DR$, $CD117$, $CD45$, $CD34$, and $CD38$), and physical properties FS (forward scatter; size of cell) and SS (side scatter; granularity of cell). We abbreviate this data set as 'flowcap'. In regards to ground truth, it is known that AML samples have a larger $CD34+$ population.

The 'pregnancy' data set [3] consists of human maternal whole-blood FCM samples. We compare 2 samples taken from each of the 18 and 10 women of the training and validation cohort during late-term pregnancy and 6 weeks postpartum. Each sample has around 300,000 cells analyzed on 13 markers ($CD123$, $CD14$, $CD16$, $CD3$, $CD4$, $CD45$, $CD45RA$, $CD56$, $CD66$, $CD7$, $CD8$, $Tbet$, $TCRgd$). Since discrepancies between subjects is a major batch effect, we further normalize for subject. For each subject, we take the calculated feature values of all of her samples and extract the difference between them and their mean.

3 Results and discussion

By incorporating the cell hierarchical structure into our feature design, we show that compared to traditional features, specific enrichment (1) finds just as or more robust significant bio-marker features associated with samples of different classes, (2) allows users to find the source cell population of change as opposed to finding all changed cell populations whether or not those are induced changes, and (3) works to isolate known and new bio-markers in real data.

3.1 Robustness

A first trait of robustness is consistency (Figure 2A). We split up our data set into two equally sized sets to see if theoretically similar data sets would yield similar BY-adjusted q-values across all cell populations. This is shown in the figure using the Spearman correlation between the two sets of q-values, and the recall, precision, and F measure over the first set — this indicates how many significant cell populations in the first set also shows up as significant in the second set. For data set pos1, the specific enrichment feature type produces scores just as high as the other features across all metrics for a significance threshold of .05.

Another robustness measure we use tests whether our feature type obtains a random uniform distribution of unadjusted p-values from our negative control data set. We show specific enrichment is able to generate unadjusted p-values in similar proximity to a uniform distribution as existing features (2B). It also yields a .05 proportion of significant features (Figure 2C) given a .05 p-value threshold.

3.2 Isolating significantly changed differential cell populations

Using our positive control data sets, we show that specific enrichment is able to isolate real non-induced significantly changed cell populations. On the cell hierarchy, specific enrichment will identify these cell populations by only flagging it and its ancestors, but not its descendants (see Section 2.4). In contrast, other features flag the real changed cell populations, their ancestors, their descendants, and if they are on a layer > 1, the ancestors and descendants of other cell populations containing some of the same markers as them on the same layer. Flagging an excess of cell populations makes it hard to interpret which cell population is most important. Also note that since we default the expected proportion value of all first layer cell populations, if there is an actual changed cell population on

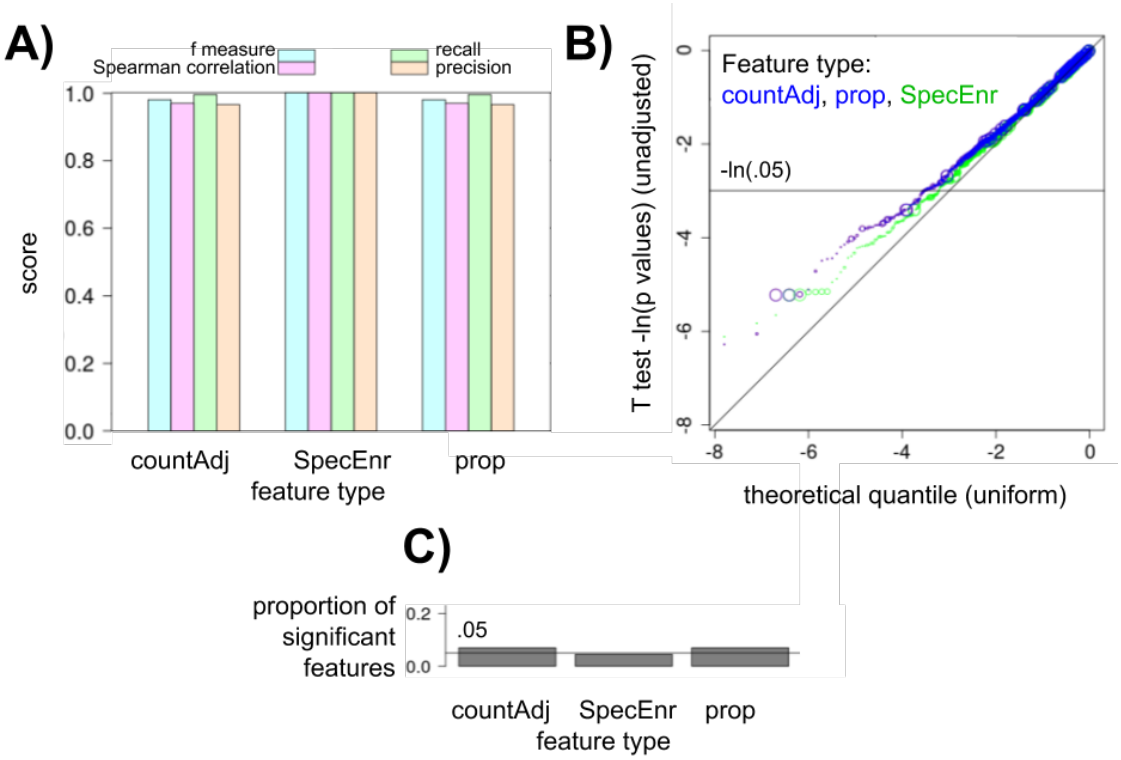

Fig. 2. Plots indicating that SpecEnr is just as robust as other flow cytometry cell population features. A) F-measure, recall, and precision scores measuring consistency between two theoretically identical data sets, B) QQ-plot for the negative control data set, and C) Proportion of cell populations with significant ($< .05$) p-values; A) is created with q-values from the pos1 data set while B) and C) are created with p-values from the ctrl data set.

the first layer (A^+ increase), we will also see its sibling cell population containing the same marker change in the reverse direction being flagged (e.g. A^- decrease).

In Figure 3, pos1 and pos3 only had a proportion change in cell populations containing marker A in the first layer and markers A and B in the second layer. While SpecEnr, proportion, and normalized count feature types flags those, the latter two also excessively flag all of their descendants as significantly differentiated. Though CytoDX does not flag all descendants, it also flags cell populations with markers other than A — notably BSpecEnr is also able to isolate only the real changed cell populations when multiple cell populations change as demonstrated on data sets pos2 and pos4 in the first and second layer respectively.

The last scenario is when we have A^+ and $A^+B^+C^+$, who are direct ancestor/descendants of each other, change simultaneously, as in pos5. As the increase in A^+ implicitly causes an increase in $A^+B^+C^+$ in addition to the latter cell populations' own change, a user can analyze $A^+B^+C^+$ on its own. To isolate A^+ or the marker condition that most influences the value of $A^+B^+C^+$, on can calculate the arc that holds the maximum value on all paths to $A^+B^+C^+$, which would be A^+.

3.3 Effectiveness on real data sets

We also demonstrate how specific enrichment can extract known significantly differentiated cell populations and isolate a range of other real changed cell populations (Figure 4). For the flowcap data set, all feature types isolate the ground truth $CD34$ marker as an important cell population defining marker. However, specific enrichment exhibits only a few important cell populations while the other features flag almost all cell populations in the cell hierarchy. On top of $CD34$, specific enrichment also directs the user down a branch containing populations defined by $CD34$ and $CD38$ revealing a more detailed story of how the change in $CD34^+$ came about — via an increase in one of its descendants.

Specific enrichment exhibits a similar contrast in behaviour in the pregnancy data set. The top 10 most significant cell populations displayed by specific enrichment leads the user down to two cell populations with markers $CD16$, $CD3$, and $TCRgd$. While prop also flags several cell populations with markers $CD16$ and $CD3$, it also associates them with markers $CD45RA$, $CD56$, and $CD66$ whom specific enrichment flagged at the top of the hierarchy separating them out as independent sources of influence on the data set.

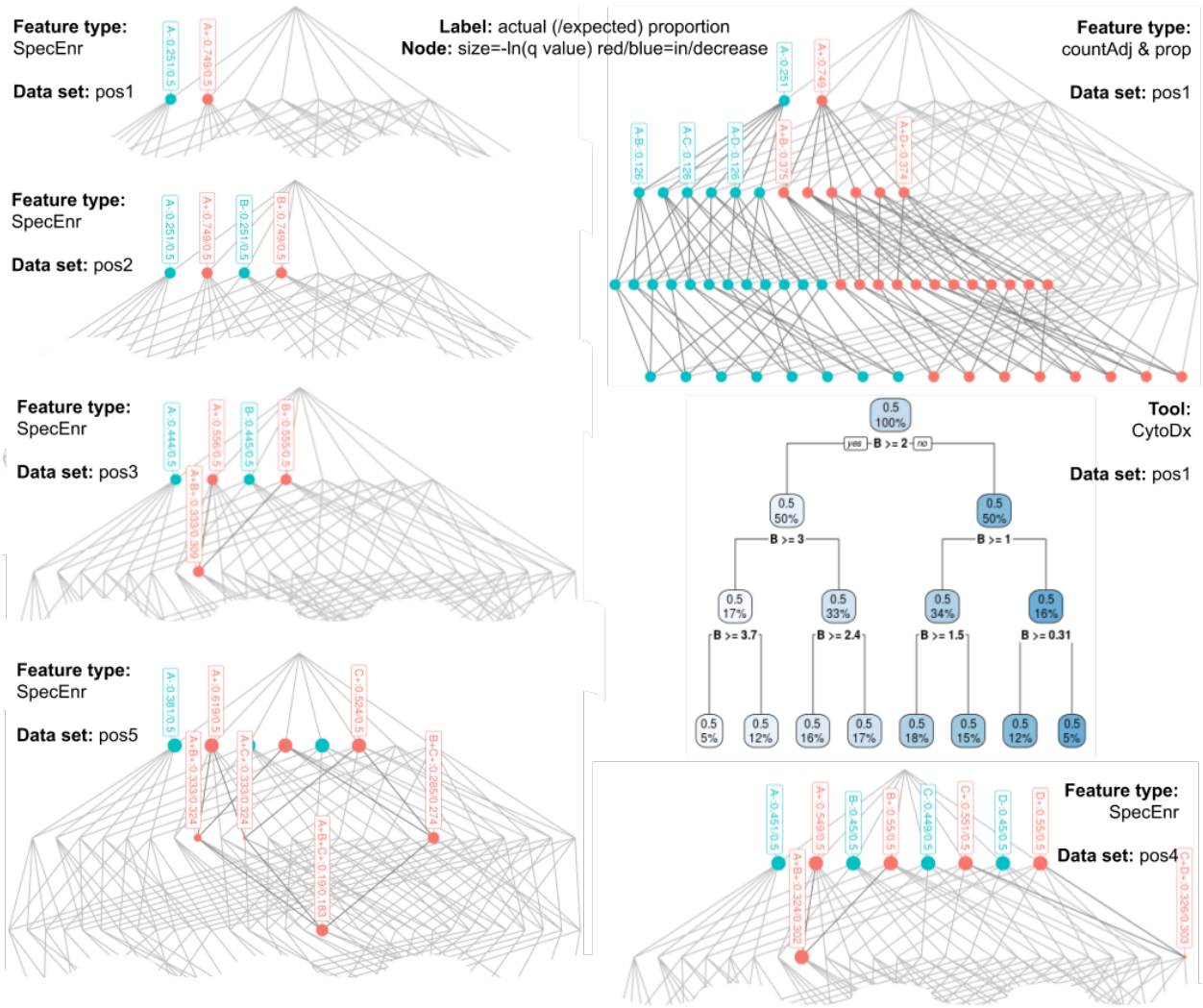

Fig. 3. SpecEnr cell hierarchy plots for data sets pos1-5, prop cell hierarchy plot for data set pos1, and a CytoDx plot for data set pos1. SpecEnr is able to only flag the real changed differential expressed cell populations.

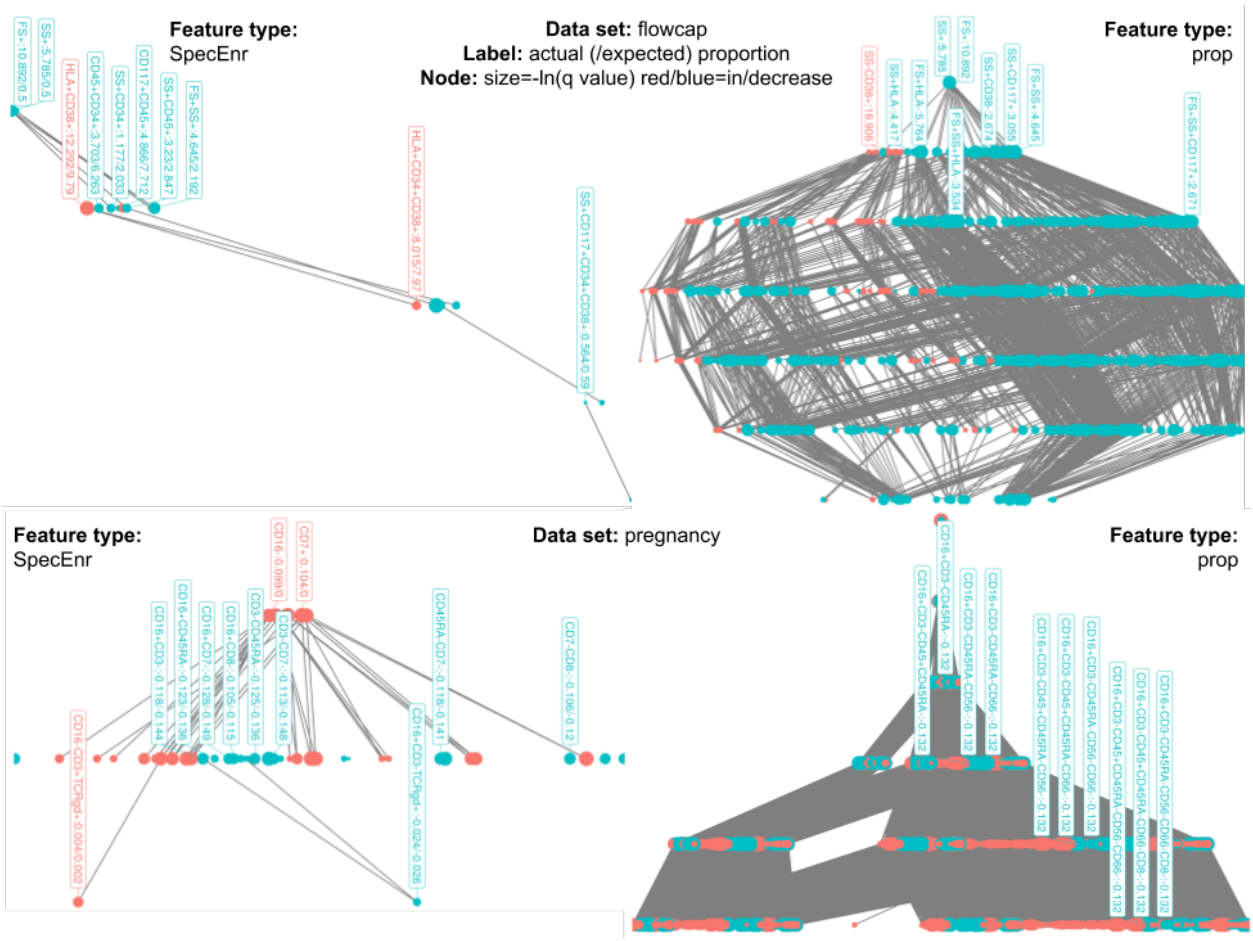

Fig. 4. SpecEnr and prop cell hierarchy plots for the flowcap and pregnancy data sets. SpecEnr is able to identify known differential cell populations and identify new ones while helping the user interpret which cell populations are most important.

4 Conclusion

In this paper, we proposed a method that can identify robust, real significantly changed cell population bio-markers between FCM samples representing phenotypes of interest. It utilizes a novel feature type, specific enrichment, to take into consideration relations between cell populations which allows for reasoned interpreted via a cell hierarchy visualization. For our future work, we will continue to analyze the behaviour of specific enrichment on different situations of cell population change and define a concrete algorithm for the user to more easily isolate the one or few best, intuitive, and explainable FCM gating strategies from the bio-markers found. Although artificial in/decreases caused by the use of proportions over normalized counts can be flagged via the cell hierarchy, we also aim to incorporate aspects of normalized count for ease of result interpretation.

Acknowledgements

We thank Albina Rahim, Nima Aghaeepour for providing the pregnancy data set.

References

1. Aghaeepour, N., Chattopadhyay, P.K., Ganesan, A., O'neill, K., Zare, H., Jalali, A., Hoos, H.H., Roederer, M., Brinkman, R.R.: Early immunologic correlates of hiv protection can be identified from computational analysis of complex multivariate t-cell flow cytometry assays. Bioinformatics **28**(7), 1009–1016 (2012)
2. Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T.R., Brinkman, R., Gottardo, R., Scheuermann, R.H., Consortium, F., Consortium, D., et al.: Critical assessment of automated flow cytometry data analysis techniques. Nature methods **10**(3), 228–238 (2013)
3. Aghaeepour, N., Ganio, E.A., Mcilwain, D., Tsai, A.S., Tingle, M., Van Gassen, S., Gaudilliere, D.K., Baca, Q., McNeil, L., Okada, R., et al.: An immune clock of human pregnancy. Science immunology **2**(15), eaan2946 (2017)
4. Azad, A., Rajwa, B., Pothen, A.: immunophenotype discovery, hierarchical organization, and template-based classification of flow cytometry samples. Frontiers in Oncology **6** (2016)
5. Benjamini, Y., Yekutieli, D., et al.: The control of the false discovery rate in multiple testing under dependency. The annals of statistics **29**(4), 1165–1188 (2001)
6. Brinkman, R.R., Aghaeepour, N., Finak, G., Gottardo, R., Mosmann, T., Scheuermann, R.H.: Automated analysis of flow cytometry data comes of age. Cytometry Part A **89**(1), 13–15 (2016)
7. Bruggner, R.V., Bodenmiller, B., Dill, D.L., Tibshirani, R.J., Nolan, G.P.: Automated identification of stratifying signatures in cellular subpopulations. Proceedings of the National Academy of Sciences **111**(26), E2770–E2777 (2014)
8. Chattopadhyay, P.K.: Toward 40+ parameter fluorescence flow cytometry. In: Cyto2014. Congress Int. Soc. Advancement Cytom (2014)
9. Chattopadhyay, P.K., Hogerkorp, C.M., Roederer, M.: A chromatic explosion: the development and future of multiparameter flow cytometry. Immunology **125**(4), 441–449 (2008)
10. Hahne, F., Gopalakrishnan, N., Khodabakhshi, A.H., Wong, C.J., Lee, K.: flowstats: Statistical methods for the analysis of flow cytometry data. R Packag. version **3**(1) (2009)
11. Hu, Z., Glicksberg, B.S., Butte, A.J.: Robust prediction of clinical outcomes using cytometry data. Bioinformatics **35**(7), 1197–1203 (2018)
12. Irish, J.M.: Beyond the age of cellular discovery. Nature immunology **15**(12), 1095–1097 (2014)
13. Kvistborg, P., Gouttefangeas, C., Aghaeepour, N., Cazaly, A., Chattopadhyay, P.K., Chan, C., Eckl, J., Finak, G., Hadrup, S.R., Maecker, H.T., et al.: Thinking outside the gate: single-cell assessments in multiple dimensions. Immunity **42**(4), 591 (2015)
14. Lin, L., Finak, G., Ushey, K., Seshadri, C., Hawn, T.R., Frahm, N., Scriba, T.J., Mahomed, H., Hanekom, W., Bart, P.A., et al.: Compass identifies t-cell subsets correlated with clinical outcomes. Nature biotechnology **33**(6), 610–616 (2015)
15. Maecker, H.T., Rinfret, A., D'Souza, P., Darden, J., Roig, E., Landry, C., Hayes, P., Birungi, J., Anzala, O., Garcia, M., et al.: Standardization of cytokine flow cytometry assays. BMC immunology **6**(1), 13 (2005)
16. Malek, M., Taghiyar, M.J., Chong, L., Finak, G., Gottardo, R., Brinkman, R.R.: flowdensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. Bioinformatics **31**(4), 606–607 (2015)
17. O'Neill, K., Aghaeepour, N., Špidlen, J., Brinkman, R.: Flow cytometry bioinformatics. PLoS computational biology **9**(12), e1003365 (2013)
18. O'Neill, K., Jalali, A., Aghaeepour, N., Hoos, H., Brinkman, R.R.: Enhanced flowtype/rchyoptimyx: a bioconductor pipeline for discovery in high-dimensional cytometry data. Bioinformatics **30**(9), 1329–1330 (2014)
19. Rahim, A., Meskas, J., Drissler, S., Yue, A., Lorenc, A., Laing, A., Saran, N., White, J., Abeler-Dörner, L., Hayday, A., et al.: High throughput automated analysis of big flow cytometry data. Methods **134**, 164–176 (2018)
20. Robinson, J.P., Rajwa, B., Patsekin, V., Davisson, V.J.: Computational analysis of high-throughput flow cytometry data. Expert opinion on drug discovery **7**(8), 679–693 (2012)
21. Robinson, M.D., Oshlack, A.: A scaling normalization method for differential expression analysis of rna-seq data. Genome biology **11**(3), R25 (2010)
22. Saeys, Y., Van Gassen, S., Lambrecht, B.N.: Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nature Reviews Immunology **16**(7), 449–462 (2016)
23. Spidlen, J., Moore, W., Parks, D., Goldberg, M., Bray, C., Bierre, P., Gorombey, P., Hyun, B., Hubbard, M., Lange, S., et al.: Data file standard for flow cytometry, version fcs 3.1. Cytometry Part A **77**(1), 97–100 (2010)
24. Tong, D.L., Ball, G.R., Pockley, A.G.: gem/gann: A multivariate computational strategy for auto-characterizing relationships between cellular and clinical phenotypes and predicting disease progression time using high-dimensional flow cytometry data. Cytometry Part A **87**(7), 616–623 (2015)
25. Van Dongen, J., Lhermitte, L., Böttcher, S., Almeida, J., Van der Velden, V., Flores-Montero, J., Rawstron, A., Asnafi, V., Lecrevisse, Q., Lucio, P., et al.: Euroflow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. Leukemia **26**(9), 1908–1975 (2012)

26. Van Gassen, S., Vens, C., Dhaene, T., Lambrecht, B.N., Saeys, Y.: Floremi: Flow density survival regression using minimal feature redundancy. Cytometry Part A **89**(1), 22–29 (2016)
27. Zare, H., Shooshtari, P., Gupta, A., Brinkman, R.R.: Data reduction for spectral clustering to analyze high throughput flow cytometry data. BMC bioinformatics **11**(1), 403 (2010)

Appendices

Appendix A

Appendix 1: Positive control results for differential cell populations that decrease in abundance

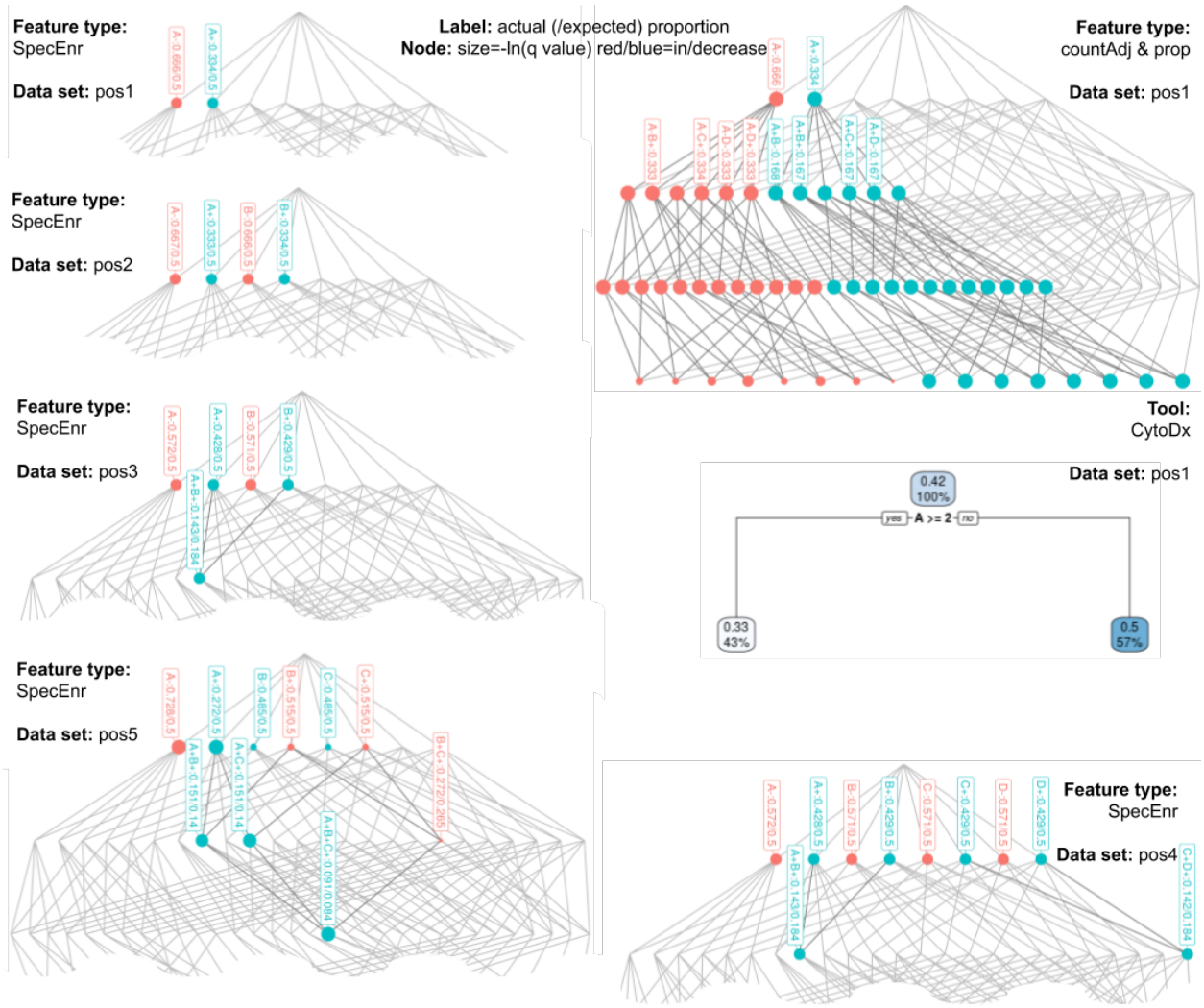

Fig. 5. SpecEnr cell hierarchy plots for data sets pos1-5, prop cell hierarchy plot for data set pos1, and a CytoDx plot for data set pos1. The positive control data sets mentioned here are the same as the ones in the paper, but with decreased abundance instead of increased abundance.

Appendix B

Appendix 2: Proof of correctness for Equation 4

Theorem 1. *Given that there is some pair of marker indices p, q s.t. $P(v^p)$ and $P(v^q)$ are independent given $P(v^{1:\ell\setminus\{p,q\}})$ and*

$$P(v^{1:\ell}) = P(v^{1:\ell\setminus p})\frac{P(v^{1:\ell\setminus q})}{P(v^{1:\ell\setminus\{p,q\}})}$$

where P is the actual proportion of cell population v defined by ℓ marker conditions, then we can identify such a p, q pair using the following

$$P(v^{1:\ell}) = \max_{p\in 1:\ell} P(v^{1:\ell\setminus p}) \min_{q\in 1:\ell\setminus p} \frac{P(v^{1:\ell\setminus q})}{P(v^{1:\ell\setminus\{p,q\}})}$$

Proof. Given our assumption, we prove our theorem for two cases in which we can find a p, q pair s.t. $P(v^p)$ and $P(v^q)$ are independent given $P(v^{1:\ell\setminus\{p,q\}})$. These are when (I) all or (II) some markers are independent of each other.

Lemma 1. *in (I), Equation 4 would always find an $p \neq q$ pair that would result in an independent pair of $P(v^p)$ and $P(v^q)$.*

Proof. Since $P(v^p)$ and $P(v^q)$ is independent for all pairs of p, q, no matter what p, q pair Equation 4 finds, $P(v^p)$ and $P(v^q)$ will be independent.

Lemma 2. *in (II), Equation 4 would always find an $p \neq q$ pair that would result in an independent pair of $P(v^p)$ and $P(v^q)$.*

Corollary 1. *we know already that (1) there must be a valid q for each p such that $P(v^p)$ and $P(v^q)$ are independent.*

Corollary 2. *We also observe that (2) there is only a single expected proportion value for $P(v^{1:\ell})$. So no matter what p is, q must be chosen such that $P(v^{1:\ell})$ is the same for all p*

Corollary 3. *Lemma 1,2 thus imply that we only need to identify one set of p, q that satisfies this and Lemma 1 to obtain a p, q pair such that $P(v^p)$ and $P(v^q)$ are independent.*

Proof. We show that Equation 4 is able to obtain a p, q pair such that $P(v^p)$ and $P(v^q)$ are independent in case (II) via proof by induction on r, $2 \leq r \leq \ell - 1$, the number of independent marker groups in a cell population as per case (II) for $\ell > 2$. An independent marker group here is defined by a group of marker indices $\mathbf{s} \subseteq \{1, ..., \ell\}$ s.t. $\mathbf{s} \neq \{\}$, the set of all marker groups combined $\{\mathbf{s}_1, ..., \mathbf{s}_r\} = \{1, ..., \ell\}$, $P(v^p)$ is independent of $P(v^q)$given $P(v^{1:\ell\setminus\{p,q\}})$ $\forall\ p \in \mathbf{s}_i, q \in \mathbf{s}_j, i \neq j, \{i, j\} \subseteq \{1, ..., r\}$, and $P(v^p)$ is dependant of $P(v^q)$ $\forall\ \{p, q\} \subseteq \mathbf{s}_i, i \in \{1, ..., r\}$.

For base case $r = 2$, our solution holds. An example on $\ell = 3$ is when A^+B^+ is a differential cell population and we are calculating the expected proportion of $A^+B^+C^+$. As C^+ is independent of both A^+ and B^+, p can be C^+ and q can be A^+ or B^+, or vice versa. We call these two marker groups containing markers that are independent of each other. If A^+ and B^+ increased at a higher

rate than C^+, p would be C^+ and q would be any of A^+, B^+ equivalent. If the reverse happened, p would be any of A^+, B^+ equivalent and q would be C^+. Note that we know the rate of change must be different therefore p and q cannot come from the same marker group. The same would happen when $\ell > 3$ in that p and q would be chosen such that they are from different marker groups.

Our inductive hypothesis states that our solution holds for all values of r up to some whole number k.

For our inductive step, we show our solution also works for $r = k + 1$. A reminder that we are proving this for case (II) where all values representing r must satisfy $2 \leq r \leq \ell - 1$. Previously, we saw a scenario where there are two independent groups of markers. If the new marker group comes from a differential expressed cell population that did not increase or decrease at a rate as high as the current maximal and minimal marker groups (e.g. A^+B^+ and C^+), our p and q remains the same and our solution holds. Otherwise, our new marker group contains marker(s), any of which can be p while q represents an existing marker (again p and q are reversible). Since we already know these marker groups are independent, this is a valid set of p, q.

As such, by the principle of mathematical induction,our solution also holds for cell populations on layers $3 \leq \ell \leq k + 2$ for some whole number k and for all $2 \leq r \leq \ell - 1$.