Open access • Posted Content • DOI:10.1101/837765

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

Alice Yue, Cedric Chauve, Maxwell W. Libbrecht, Ryan R. Brinkman

Institutions: Simon Fraser University, BC Cancer Agency

Published on: 16 Nov 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Population

Automated identification of maximal differential cell populations in flow cytometry data

Alice Yue[1]*, Cedric Chauve[2,3]†‡, Maxwell Libbrecht[1]§§ and Ryan R. Brinkman[4,5]§¶

[1]Computing Science, 9971 Applied Sciences Building, Simon Fraser University, 8888 University Drive, Burnaby, V5A 1S6, BC, Canada.
[2]Mathematics, Simon Fraser University, Burnaby, V5A 1S6, BC, Canada.
[3]LaBRI, University of Bordeaux, Bordeaux, 33076, Nouvelle-Aquitaine, France
[4]Terry Fox Laboratory, 13th floor, BC Cancer Research Centre, BC Cancer Agency, 675 West 10th Avenue, Vancouver, V5Z 1L3, BC, Canada.
[5]Medical Genetics, University of British Columbia, Vancouver, V6T 1Z4, BC, Canada

Running head: Finding maximal differential cell populations

Funding statement: This work was funded by Simon Fraser University, the Natural Sciences and Engineering Research Council of Canada, and the National Institutes of Health (Award ID: 5R01GM118417-04).

Data availability statement: The flowGraph package, scripts, and data used in this paper are available at https://github.com/aya49/flowGraph; the package is submitted to BioConductor, has passed all tests and is under review.

Conflict of interest disclosure: RR Brinkman serves on the advisory board of Cytapex Bioinformatics Inc.

*E-mail: aya43@sfu.ca; ORCID ID: 0000-0002-8296-8941
†Corresponding author.
‡E-mail: cedric.chauve@sfu.ca; ORCID ID: 0000-0001-9837-1878
§E-mail: maxwl@sfu.ca; ORCID ID: 0000-0003-2502-0262
¶E-mail: rbrinkman@bccrc.ca; ORCID ID: 0000-0002-9765-2990

Abstract

We introduce a new cell population score called SpecEnr (specific enrichment) and describe a method that discovers robust and accurate candidate biomarkers from flow cytometry data. Our approach identifies a new class of candidate biomarkers we define as driver cell populations, whose abundance is associated with a sample class (e.g. disease), but not as a result of a change in a related population. We show that the driver cell populations we find are also easily interpretable using a lattice-based visualization tool. Our method is implemented in the R package flowGraph, freely available on GitHub (github.com/aya49/flowGraph) and will be available BioConductor.

Key words: Automated analysis, Statistical analysis, Exploratory data analysis, Flow cytometry, Bioinformatics

Introduction

A major goal in flow cytometry (FCM) analysis is the identification of candidate biomarkers. The most common candidates are differential cell populations (DCPs). These are cell populations whose proportional abundances (i.e., the relative quantity of cells in a cell population) differ significantly between samples of separate classes (e.g. disease vs healthy). Commonly used metrics for proportional abundance are cells per μL of blood and proportion (i.e, the ratio between the count of cells in a population and some parent population).

Here, we propose the concept of *maximal differential cell populations* (MDCPs). MDCPs are DCPs whose change in proportional abundance is only significantly associated with its sample class, as opposed to being the result of a proportional abundance change in a related DCP. For example, if there is a significant decrease in the proportion of helper T-cells in samples from sick individuals compared to those from healthy individuals, then helper T-cells is a DCP. However, if the proportional abundance of all types of T-cells decrease at a similar rate, then we can hypothesize that the disease decreases the proportional abundance of T-cells. It follows that T-cells and all of its child populations, including helper T-cells, are DCPs but only T-cells is a MDCP. MDCPs are preferable candidate biomarkers because their proportional abundance change is only driven by their association with a sample class. We refer to such cell populations as driver cell populations. To our knowledge, while there are many methods that find biomarker candidates by identifying DCPs, there are no methods that do so by isolating only the MDCPs among those DCPs.

Most methods identify DCPs either as a byproduct of another procedure or by evaluating a limited set of prespecified cell populations [2, 4–11, 13, 15, 19]. The latter group of methods compare the proportional abundance of each cell population across samples using some statistical significance test. Comparing proportional abundance of cell populations across samples is effective for finding DCPs but not MDCPs as they do not account for the relationship between cell populations.

To address these shortcomings, we find MDCPs by comparing the SpecEnr of cell populations across samples. SpecEnr is a novel cell population score, a numerical metric that is derived from the proportional abundance metric, proportion, and accounts for the relationship between cell populations. In this paper, we: 1) Define and formulate the

problem of finding driver cell populations by identifying MDCPs; 2) Introduce a cell population score, SpecEnr (specific enrichment), that accounts for dependencies between parent and child cell populations; 3) Describe a method that harnesses SpecEnr properties to find robust, accurate, and easily interpretable driver cell populations. We hypothesize that identifying MDCPs will aid the understanding of disease etiology.

Methods and Materials

Cell hierarchy

[Figure 1 about here.]

To visualize the relationship between cell populations, we use the cell population hierarchy of a sample. A *cell population hierarchy* or cell hierarchy for short, is a directed acyclic graph where nodes represent cell populations and arcs represent the relationship between cell populations (Figure 1). We define a cell population as a set of cells with similar fluorescent intensity (FI) values for a set of $0 \leq \ell \leq L$ measurements (e.g. markers, SSC, FSC). For simplicity, we define a measurement condition as a combination of a measurement and a positive$^+$ or negative$^-$ expression indicator. For example, A^+B^- contains two measurement conditions (A^+ and B^-) and represents a cell population whose cells have FI greater and less than the given thresholds for measurements A and B respectively. Note our method is applicable for multiple thresholds per measurement (Supplementary Material). We define the ℓ'th layer of the cell hierarchy as the set of all nodes whose label contains exactly ℓ unique measurement conditions. It then follows that a cell hierarchy has $L+1$ possible layers. The 0'th layer contains the root cell population comprising all cells. Each arc points from a 'parent' cell population to its 'child' sub-population defined by the addition of one measurement condition. For example, if there are three measurements $\{A, B, C\}$, then there are arcs from the node representing the cell population labelled A^+ to the nodes labelled A^+B^+, A^+B^-, A^+C^+, and A^+C^-.

Preprocessing

Our approach takes as input a vector of cell population proportions for each FCM sample generated using any any suitable manual or automated approach. If users want to analyze a cell population defined by more than two measurement conditions, we require that this vector also contain all of this cell populations' parent and grandparent cell populations as defined on the cell hierarchy introduced in the previous section. For our experiments, given a FCM sample containing a cell $\times$ measurement matrix and threshold gates obtained via gating, we use flowType [9] to identify all possible cell populations and enumerate their cell count. Next we normalize cell counts with respect to the total cell count by converting counts into proportions by taking the cell count of each cell population over the total number of cells in the sample.

Cell population score: SpecEnr

To obtain SpecEnr, we compare the actual proportion of a cell population with its expected proportion: the proportion we expect a cell population to have given the proportion of its ancestors. By doing so, we can evaluate its proportion changes independent of the affects incurred by its ancestors.

Expected proportion

We denote the actual proportion P of any node $v^{1:\ell}$ in layer ℓ by $P(v^{1:\ell})$ where, $1{:}\ell$ are the indices of the measurement conditions its label contains. For example, cell population $A^+B^+C^-$ has three measurement conditions and can therefore be denoted as $v^{1:3}$; subsequently, we can denote its parents A^+C^- and A^+B^+ as $v^{\{1:3\}\setminus 2}$ and $v^{\{1:3\}\setminus 3}$ by excluding the second and third measurement conditions.

We denote the expected proportion $P'(v^{1:\ell})$ of a cell population v as its proportion if it satisfies an assumption: let's assume that $P(v^1)$ (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 this assumption to any p, q pair, $p \in 1{:}\ell$ and $q \in 1{:}\ell \setminus p$, we get

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

Our assumption requires $P(v^{1:\ell\setminus\{p,q\}})$ to exist. Therefore, expected proportion is only calculated for cell populations in layers $\ell \geq 2$. For the root node, we initialize its expected proportion as 1. For the nodes in layer one, we initialize their expected proportions to .5.

In Equation 3, we assume all measurement condition pairs q, p should be independent of each other. Now let's suppose this assumption does not hold for cell population A^+C^-. While A^+ and C^- are dependent on each other, B^+ is independent of both A^+ and C^-. In this case, the assumption we made in Equation 2 only holds for cell population $A^+B^+C^-$ when $q \in \{1, 2\}$ and $p = 3$. We do not want to flag $A^+B^+C^-$ as maximally differential as its proportion change is completely dependent on cell populations A^+C^- and B^+. Therefore, we relax our assumption in Equation 3 to: there must be some p, q

pair such that $P(v^p)$ is independent of $P(v^q)$. If so, then $P(v^{1:\ell})$ can be calculated as follows.

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

In the context of MDCP, if $P(v) = P'(v)$, then v is not a MDCP. This is because v's proportion change can be attributed to its ancestor cell populations and it is therefore not maximally differential.

Additional details on algorithmic & runtime and proof of correctness are provided in the Supplementary Material.

SpecEnr

Given the expected proportion of cell population v calculated using Equation 4, we can get SpecEnr by taking the natural log of the its actual proportion P over its expected proportions P'.

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

SpecEnr accounts for the dependency of a cell population on its ancestors. For example, if a cell population has a SpecEnr value of 0, then its proportional abundance is completely dependent on that of its ancestors. Otherwise, it contains measurement conditions that are all dependent on each other, where $P(v^p)$ is dependent on $P(v^q)$ for all $\{p, q\} \in 1{:}\ell$ (i.e. Equation 3 does not hold for any p, q).

Maximal differential cell populations

A MDCP must $v^{1:\ell}$ satisfy two conditions.

1. A MDCPs' SpecEnr must be significantly different between samples according to a filtered adjusted T-test we describe in the next section.

2. A MDCP must also be maximal, in that it must not have any direct descendants who meet our first condition.

The second condition is required because our first is also satisfied by direct ancestors of a MDCP as its ancestor cell populations are defined by a subset of measurement conditions defining the MDCP.

Significance test

To test if a cell population satisfies our first condition for MDCPs (i.e. its SpecEnr is significantly different across samples), we apply the t-test on SpecEnr values for each cell population across two sets of sample from (e.g. the control and experiment group). We adjust these p-values ρ_v for each cell population v using layer-stratified Bonferroni correction to obtain our final adjusted p-values ρ'_v. We do so by multiplying our p-values with the number of cell populations in the layer on which cell population v resides m_ℓ and the total number of layers $L + 1$ (including the layer 0; see Supplementary Material for additional details). For example, if we are working with four measurements *A*, *B*, *C*, and *D*, we multiply the p-value of A^+ by 8 and 5, the number of nodes in layer one and the total number of layers.

$$\rho'_v = \rho_v \cdot m_\ell \cdot (L + 1)$$

Note that users can use any significance test and p-value adjustment strategy they deem suitable for their experiment. We use a p-value threshold $< .05$ to determine if a cell population p-value is significant and potentially maximally differential.

Filters

In some cases, the p-value obtained by evaluating SpecEnr may be falsely significant when dealing with small or noisy data sets. As a cell populations' proportion gets close to 0, the actual vs expected proportion ratio used to calculate SpecEnr becomes inflated. As well, if we are conducting significance tests on cell populations with SpecEnr values of 0 (i.e. actual and expected proportions are the same) model-based significance tests (e.g. T-test) are highly influenced by outliers and rank-based significant tests (e.g. Wilcoxan) are influenced by random ordering of 0's. To ensure our SpecEnr p-values are valid, we mark cell populations as insignificant if they do not 1) have a mean count of > 50 events to prevent inflated ratios, 2) have significantly different actual vs expected proportions for at least one of the sample classes, and 3) contain actual and expected proportions that are different at different rates across both sample classes. See Supplementary Materials for details. Note that we use a significance threshold of $< .05$ for all t-test p-values on filter related significance tests. We show an example of these filters in the Supplementary Material. For brevity, we call the p-values obtained using SpecEnr and proportion, SpecEnr p-values and proportion p-values respectively.

Experiment data

To confirm that our approach is able to identify known MDCPs we prepared synthetic negative and positive control data sets and used two previously published biological data sets.

Synthetic data

- **neg1** (Negative control): For each cell, we assigned it to be positive$^+$ for each measurement with a 50% probability.
  - Samples: 10 control vs 10 experiment (300,000 cells/sample).
  - Measurements: A, B, C, and D.
- **pos1** (Positive control 1): Same as neg1, except in the experiment samples, cell population A^+ is increased by 50%. More specifically, in each $R \times L$ matrix, we duplicated a random sample of half the rows with a measurement A FI higher than our given threshold gate.
- **pos2** (Positive control 2): Same as pos1, except instead of A^+, $A^+B^+C^+$ is increased by 50%.
- **pos3** (Positive control 3): Same as pos1, except instead of A^+, A^+B^+ and D^+ are both increased by 50% causing a unique increase in cell population $A^+B^+D^+$.

Biological data

- **flowcap** (FlowCAP-II AML data set): This data set is from the FlowCAP-II [15], AML challenge, panel 6. It is known that AML samples have a larger $CD34^+$ population [15].
  - Samples: 316 healthy vs 43 AML positive subjects' blood or bone marrow tissue samples ($\sim$60,000 cells/sample).
  - Measurements: *HLA-DR*, *CD117*, *CD45*, *CD34*, and *CD38*.
- **pregnancy** (Immune clock of pregnancy data set): So far, there has been no experiments that identified ground truth driver cell populations for the pregnancy data set [13]. However, the original authors were able to train classifiers on the same patients using FCM and multi-omics data [14]. Therefore, we hypothesize that we

will be able to find MDCPs in this data set that are associated with the sample classes listed below.

- Samples: 28 late-term pregnancy vs 28 6-weeks postpartum human maternal whole-blood samples (~300,000 cells/sample); Samples are taken from each of the 18 and 10 women of the training and validation cohort during late-term pregnancy and 6 weeks postpartum.
- Measurements: *CD123*, *CD14*, *CD16*, *CD3*, *CD4*, *CD45*, *CD45RA*, *CD56*, *CD66*, *CD7*, *CD8*, *Tbet*, and *TCRgd*.
- To account for possible batch effects associated with the subjects who provided the FCM samples, we used the paired t-test with respect to subject.

Results

SpecEnr p-values are robust. In this experiment, we hypothesized that theoretically similar data sets yield similar unadjusted p-values across all cell populations. When we compared the unadjusted SpecEnr p-values across these two data sets using the Spearman correlation, we obtained a perfect score of 1. We saw the same result with metrics recall, precision, and F measure over the first set. These results indicate that significant cell populations in the first set also show up as significant in the second set. SpecEnr p-values are also statistically sound [16]. Using SpecEnr, we were able to generate a random uniform distribution of unadjusted p-values on our negative control data set neg1. It follows that 5% of the SpecEnr p-values were below our .05 threshold (See Supplementary Material for added detail).

SpecEnr p-values help identify accurate driver cell populations in synthetic data sets. pos1 and pos2's driver cell populations were A^+ and $A^+B^+C^+$ (Figure 2). While both SpecEnr and proportion p-values flagged these cell populations, when we observed SpecEnr p-values the descendants of these driver cell populations were not flagged as significant. Therefore, we could quickly identify A^+ and $A^+B^+C^+$ as driver cell populations as intended. This was also true when multiple driver cell populations were present in lower layers of the cell hierarchy. In pos3, where both A^+B^+ and D^+ were increased to cause a unique change in $A^+B^+D^+$; we saw that SpecEnr p-values were only significant for those cell populations and their ancestors. Results from our positive control data sets were equivalent when the same cell populations decreased instead of increased in proportional abundance (Supplementary Material).

SpecEnr p-values flag known and novel driver cell populations in real data sets. For the FlowCAP data set, SpecEnr directs users down a branch of the cell hierarchy from physical properties SS^+ and FS^- to $FS^-SS^+CD117^+45^+$ and $HLA^+CD117^-CD45^+CD34^+$. While *HLA* and *CD117* are variably expressed on cells in FCM samples from subjects with AML [17, 18], *CD34* and *CD45* are expressed on blast cells [1, 3]. This is important as the abundance of blast cells aid in diagnosis of AML [15].

In the pregnancy data set, the top most significant cell populations displayed by our statistical test shows an up-regulation in cell populations containing *CD3*, *CD45*, and

CD45RA (e.g. $CD3^{+}CD45RA^{+}CD56^{-}Tbet^{-}$). SpecEnr p-values also indicate that cell populations containing measurements *CD8* and *CD16* are significantly down-regulated. Meanwhile, proportion p-values flag all DCPs in the cell hierarchy as significant.

[Figure 2 about here.]

Discussion

In this paper, we introduced a new cell population score, SpecEnr, and a method that integrates SpecEnr to identify MDCPs. We showed that the results of our method is statistically sound, accurate, and easily interpretable.

In the FlowCAP-II challenge, the AML data set was used to evaluate how well methods are able to classify samples belonging to healthy and AML positive subjects. Among the competing methods, those that used cell population proportions for classification were DREAM–D, flowCore/flowStats, flowPeakssvm/Kmeanssvm, flowType/FeaLect, PBSC, BCB/SPADE, SWIFT. All of these methods assume that cell count and proportion may be used to differentiate between the two classes of samples. However, cell count and proportion do not account for relations between cell populations, making it difficult to isolate the MDCPs among the DCPs (Figure 2). To account for these relationship, one can manually analyze the ratio of the count of cells in a population over all of its direct parent populations. However, given L measurements, there are $3^L \cdot \frac{2L}{3}$ such relationships not including the relationship between a cell population and its indirect ancestors [9]. In contrast to comparing 3^L cell population scores, directly comparing cell population relations becomes computationally impractical.

SpecEnr mitigates both challenges as it is a cell population score that accounts for relations between cell populations. Its p-values isolated only the few ground truth driver cell populations (MDCP e.g. $SS^{-}CD34^{+}$). Hence, our results not only reveal known driver cell population $CD34^{+}$ but also provide visualizations signifying that their change may have come about because of a change in its descendants exposing novel driver cell populations.

We also observed this contrast in behaviour between SpecEnr and proportion p-values in the pregnancy data set. Our hypothesis for this data set was that we should be able to find MDCPs because [13] were able to use L_1, L_2, and cell signal pathway regularized regression to classify samples taken from women at different stages of pregnancy. The original authors used the same assumption in [4] which implies that there exists MDCPs in the pregnancy data set. However, because these methods find candidate biomarkers as a byproduct of a sample classification method, there was no way of verifying whether the candidate biomarkers they inferred are simply DCPs or are also MDCPs. Our method answers this question by providing users a way to differentiate between the two while

verifying our hypothesis validating the existence of MDCPs in the pregnancy data set.

Since SpecEnr is calculated using proportions, it is prone to the same issue that occur when using proportions directly. That is, changes in proportion of cell populations must sum to 0 analogous to a zero-sum game. For example, in pos1, A^+'s abundance doubled, so its proportion increased from .5 to .66; but A^-'s proportion decreased from .5 to .33. More generally, if a cell population is differential, it will induce a change in the proportion of all cell populations that are labelled using the same set of measurements as it; because these cell populations are mutually exclusive. Another example of this are the $\{A^{\{+,-\}}B^{\{+,-\}}C^{\{+,-\}}\}$ cell populations from pos2. If the driver cell population resides in layers > 1, then it is easily identifiable as the cell population with the largest magnitude of change. However, this is something we would like to account for in future work such that we only flag the driver cell populations and not the cell populations it affects in the context of proportions.

We mentioned that the SpecEnr p-values need to be further filtered to mitigate the false positive results that may occur because of noise or lack of data. While the filters work in practice, it would be best if SpecEnr itself provide more reliable p-values. For example, instead of a ratio, there may be better ways of comparing actual vs expected proportions.

Finally, we claimed that a t-test on SpecEnr will yield a significant p-value on driver cell populations and their ancestors. While this makes driver cell populations intuitive to find on a cell hierarchy plot, ideally, we should only flag the driver cell populations as significant and not their ancestors. By preventing excessive flagging of ancestor populations, we open the door for more expressive and detailed anecdotes in results interpretation.

Acknowledgements

This work was funded by Simon Fraser University, the Natural Sciences and Engineering Research Council, and the National Institutes of Health.

References

[1] Quek L, Otto GW, Garnett C, Lhermitte L, Karamitros D, Stoilova B, et al. Genetically distinct leukemic stem cells in human CD34- acute myeloid leukemia are arrested at a hemopoietic precursor-like stage. Journal of Experimental Medicine. 2016;213(8):1513–1535.

[2] Cossarizza A, Chang HD, Radbruch A, Akdis M, Andrä I, Annunziato F, et al. Guidelines for the use of flow cytometry and cell sorting in immunological studies. European journal of immunology. 2017;47(10):1584–1797.

[3] Harrington AM, Olteanu H, Kroft SH. A dissection of the CD45/side scatter "blast gate". American journal of clinical pathology. 2012;137(5):800–804.

[4] Hu Z, Glicksberg BS, Butte AJ. Robust prediction of clinical outcomes using cytometry data. Bioinformatics. 2018;35(7):1197–1203.

[5] Azad A, Rajwa B, Pothen A. immunophenotype Discovery, hierarchical Organization, and Template-Based classification of Flow cytometry samples. Frontiers in Oncology. 2016;6.

[6] Zare H, Shooshtari P, Gupta A, Brinkman RR. Data reduction for spectral clustering to analyze high throughput flow cytometry data. BMC bioinformatics. 2010;11(1):403.

[7] Van Gassen S, Vens C, Dhaene T, Lambrecht BN, Saeys Y. FloReMi: Flow density survival regression using minimal feature redundancy. Cytometry Part A. 2016;89(1):22–29.

[8] Tong DL, Ball GR, Pockley AG. 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. 2015;87(7):616–623.

[9] O'Neill K, Jalali A, Aghaeepour N, Hoos H, Brinkman RR. Enhanced flowType/RchyOptimyx: a Bioconductor pipeline for discovery in high-dimensional cytometry data. Bioinformatics. 2014;30(9):1329–1330.

[10] Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. Proceedings of the National Academy of Sciences. 2014;111(26):E2770–e2777.

[11] Lin L, Finak G, Ushey K, Seshadri C, Hawn TR, Frahm N, et al. COMPASS identifies T-cell subsets correlated with clinical outcomes. Nature biotechnology. 2015;33(6):610–616.

[12] Rahim A, Meskas J, Drissler S, Yue A, Lorenc A, Laing A, et al. High throughput automated analysis of big flow cytometry data. Methods. 2018;134:164–176.

[13] Aghaeepour N, Ganio EA, Mcilwain D, Tsai AS, Tingle M, Van Gassen S, et al. An immune clock of human pregnancy. Science immunology. 2017;2(15):eaan2946.

[14] Peterson LS, Stelzer IA, Tsai AS, Ghaemi MS, Han X, Ando K, et al. Multiomic immune clockworks of pregnancy. In: Seminars in Immunopathology. Springer; 2020. p. 1–16.

[15] Aghaeepour N, Finak G, Hoos H, Mosmann TR, Brinkman R, Gottardo R, et al. Critical assessment of automated flow cytometry data analysis techniques. Nature methods. 2013;10(3):228–238.

[16] Ware JH, Mosteller F, Delgado F, Donnelly C, Ingelfinger JA. P values. Medical uses of statistics. 1986;2:181–200.

[17] Wetzler M, McElwain B, Stewart C, Blumenson L, Mortazavi A, Ford L, et al. HLA-DR antigen-negative acute myeloid leukemia. Leukemia. 2003;17(4):707–715.

[18] Pomerantz A, Rodríguez-Rodríguez S, Demichelis-Gómez R, Barrera-Lumbreras G, Barrales-Benítez OV, Díaz-Huízar MJ, et al. Importance of CD117 in the Assignation of a Myeloid Lineage in Acute Leukemias. Archives of Medical Research. 2017;48(2):212–215.

[19] Chen Y, Calvert RD, Azad A, Rajwa B, Fleet J, Ratliff T, et al. Phenotyping Immune Cells in Tumor and Healthy Tissue Using Flow Cytometry Data. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; 2018. p. 73–78.

List of Figures

1 An example of a cell population hierarchy representation of a FCM sample and its cell populations defined by measurements A, B, and C. 19
2 Cell hierarchy plots for synthetic data sets pos1-3 and real data sets flowcap and pregnancy. Only significant cell population nodes are emphasized. 20

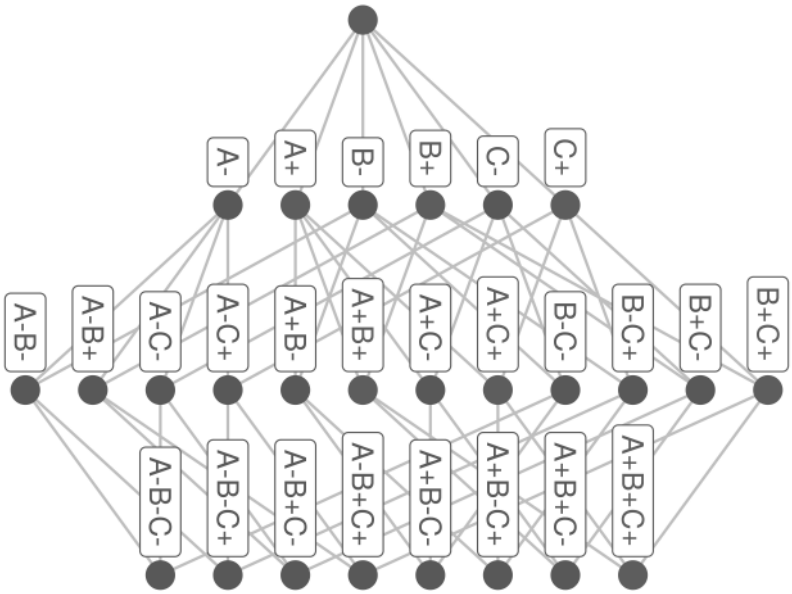

FIGURE 1: An example of a cell population hierarchy representation of a FCM sample and its cell populations defined by measurements A, B, and C.

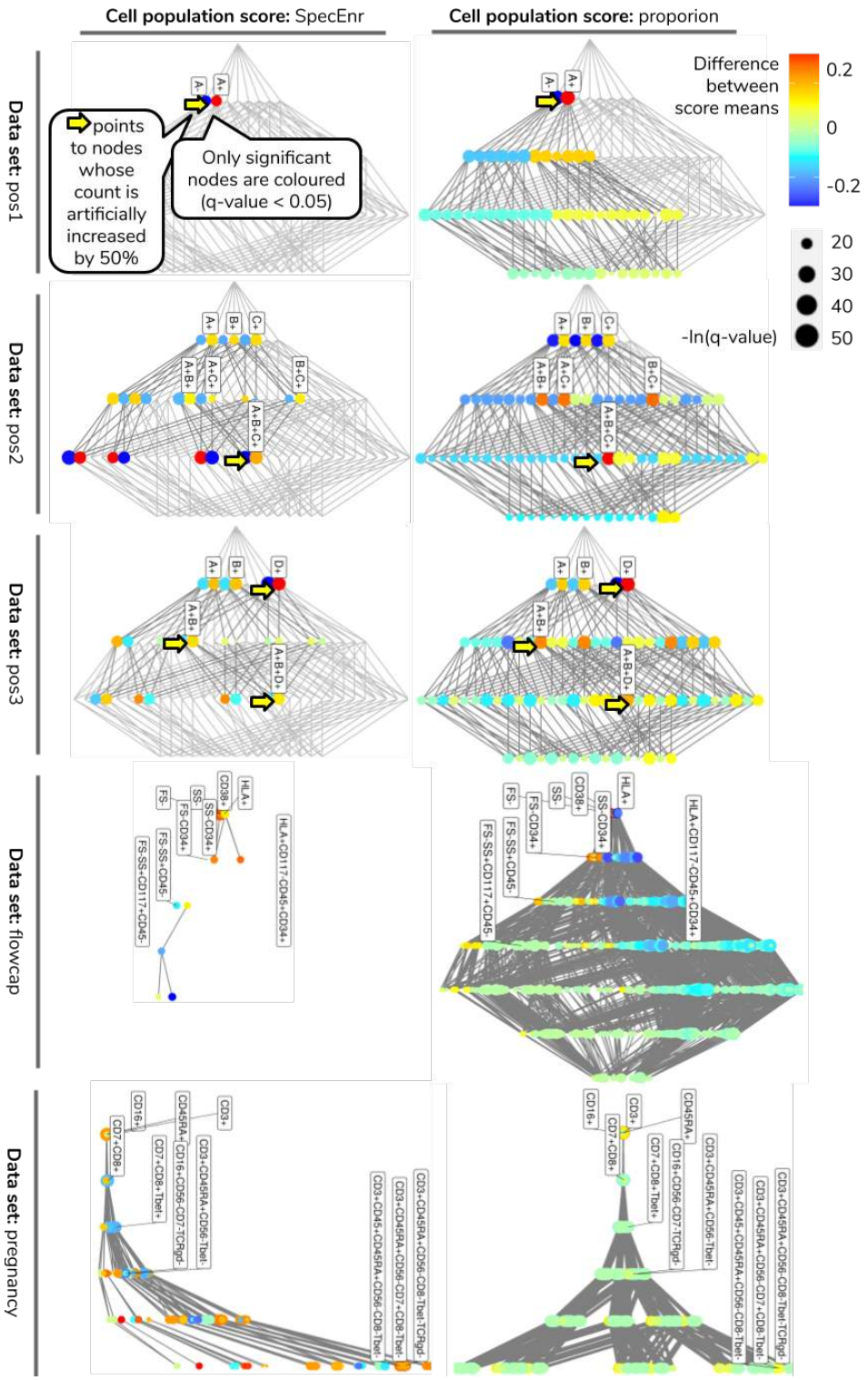

FIGURE 2: Cell hierarchy plots for synthetic data sets pos1-3 and real data sets flowcap and pregnancy. Only significant cell population nodes are emphasized.