ORIGINAL ARTICLE

Cytometry PART A
Journal of the International Society for Advancement of Cytometry

MIFlowCyt

RchyOptimyx: Cellular Hierarchy Optimization for Flow Cytometry

Nima Aghaeepour,[1] Adrin Jalali,[1] Kieran O'Neill,[1] Pratip K. Chattopadhyay,[2] Mario Roederer,[2] Holger H. Hoos,[3] Ryan R. Brinkman[1,4]*

[1]Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, British Columbia, Canada

[2]Vaccine Research Center, National Institute of Health, Bethesda, Massachusetts

[3]Department of Computer Science, University of British Columbia, British Columbia, Canada

[4]Department of Medical Genetics, University of British Columbia, British Columbia, Canada

Received 15 May 2012; Revision Received 7 August 2012; Accepted 5 September 2012

Additional Supporting Information may be found in the online version of this article.

Nima Aghaeepour and Adrin Jalali contributed equally to this work.

Grant sponsor: ISAC Scholar Awards (NA and PKC); Grant sponsor: MSFHR/CIHR scholarship to NA; Grant sponsor: University of British Columbia's 4YF scholarship to NA; Grant sponsor: Michael Smith Foundation for Health Research Scholar Award to RRB; Grant sponsor: NIH/NIBIB; Grant number: R01 EB008400; Grant sponsor: NSERC Discovery Grant held by HH; Grant sponsor: Canadian Cancer Society; Grant number: 700374; Grant sponsor: The Terry Fox Foundation; Grant sponsor: The Terry Fox Research Institute; Grant sponsor: NIAID Intramural Research Program;

• Abstract

Analysis of high-dimensional flow cytometry datasets can reveal novel cell populations with poorly understood biology. Following discovery, characterization of these populations in terms of the critical markers involved is an important step, as this can help to both better understand the biology of these populations and aid in designing simpler marker panels to identify them on simpler instruments and with fewer reagents (i.e., in resource poor or highly regulated clinical settings). However, current tools to design panels based on the biological characteristics of the target cell populations work exclusively based on technical parameters (e.g., instrument configurations, spectral overlap, and reagent availability). To address this shortcoming, we developed RchyOptimyx (cellular hieraRCHY OPTIMization), a computational tool that constructs cellular hierarchies by combining automated gating with dynamic programming and graph theory to provide the best gating strategies to identify a target population to a desired level of purity or correlation with a clinical outcome, using the simplest possible marker panels. RchyOptimyx can assess and graphically present the trade-offs between marker choice and population specificity in high-dimensional flow or mass cytometry datasets. We present three proof-of-concept use cases for RchyOptimyx that involve 1) designing a panel of surface markers for identification of rare populations that are primarily characterized using their intracellular signature; 2) simplifying the gating strategy for identification of a target cell population; 3) identification of a non-redundant marker set to identify a target cell population.

• Key terms

polychromatic flow cytometry; mass cytometry; exploratory data analysis; cellular hierarchy; graph theory; gating; marker panel; bioinformatics; statistics

RECENT advances in FCM instrumentation and reagents have enabled high-dimensional analyses to identify large numbers of cell populations with potentially significant correlations to an external outcome. However, studies often fail to characterize the complex relationships between the markers involved in the identification of these cell populations. Revealing this information can provide additional insight into the biological characteristics of the populations identified. The choice of markers for new panels has been a source of ongoing debate, including efforts such as the Human ImmunoPhenotyping Consortium (HIPC), the Federation of Clinical Immunology Societies Federation of Clinical Immunology Societies (FOCiS) sponsored Flow Immunophenotyping Technical Meetings (FITMaN), and the Optimized Multicolor Immunophenotyping Panels (OMIPs) articles (1–12). Understanding the relationships between the markers involved in identification of the target cell population and the characteristics of that cell population (e.g., its correlation with a clinical outcome) is fundamental to the design of effective marker panels. For example, one could use a high-dimensional flow or mass cytometry assay to measure a large list of candidate markers. However, this can result in parsing the cells into (e.g., clinically) redundant subsets (13). Excluding these redundancies (e.g., markers less important

ISAC
International Society for Advancement of Cytometry

*Correspondence to: Ryan R. Brinkman, Terry Fox Laboratory, BC Cancer Research Centre, 675 West 10th Avenue, Vancouver, BC, V5Z 1L3, Canada

Email: rbrinkman@bccrc.ca

Published online 8 October 2012 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/cyto.a.22209

for prediction of a clinical outcome) will result in a panel of the most clinically relevant markers.

High-dimensional FCM data is usually analyzed using a laborious sequential manual analysis procedure in which a series of thresholds or two-dimensional polygons (or gates) are applied to histograms or scatter plots of markers [e.g., (14,15)]. However, manual gates provide little insight into the relative importance of each gate to the final results. For example, consider a six color assay with markers named 1–6. If the expression of each marker is considered to be on, off, or does not matter (e.g., markers named 1, 2, and 3 in phenotype 1^+2^-, respectively), a total of $3^6 = 729$ cell populations can be distinguished based on these markers. A given immunophenotype involving all six of these markers (e.g., $1^+2^-3^+4^-5^+6^-$) can have $2^6 = 64$ parent populations (e.g., 1^+, 1^+2^-). Quantifying the relationship between the cell population of interest and these parent populations is fundamental to our understanding of the importance of the markers for different gating strategies. The order in which the gates are applied to the data is not important, as long as all of the gates are used (i.e., sequential gating is commutative). However, to decrease the size of the marker panel, the relative importance of the gates should be determined. For example, the measurement of the phenotype mentioned above using only five colors requires the determination of the importance of each marker to identify and remove the least important one (i.e., the identification of the parent population with five markers that is most similar to the original phenotype). This is further complicated by the fact that some cell populations can be identified using more than one combination of markers and gating strategy; therefore, each marker can be used in different positions in the gating hierarchy and can have different priorities, depending on the choice of the gating strategy. For example, the 3^+ gate is involved in both $1^+2^-3^+$ and $3^+4^-5^+$, both parents of the $1^+2^-3^+4^-5^+6^-$ phenotype described above. However, depending on the amount of redundancy between marker 3 and others, this marker can have different levels of importance for these two parent populations.

Another use-case for measuring the importance of the markers is the investigation of a large number of closely related phenotypes (e.g., those identified by bioinformatics pipelines) by identifying their common parent populations. Several computational tools have been developed for automated identification of cell populations [e.g., (16–26)] and recent studies have used these tools to identify novel cell populations that correlate with clinical outcomes [e.g., (27–31)]. In addition, the results of the FlowCAP-II project[1] have shown that several algorithms can accurately and reproducibly identify cell populations correlated with external outcomes. However, these algorithms provide limited information regarding the importance of the markers involved in defining the cell populations (27,32). This situation is even more complicated than sequential manual gating, since most of these bioinformatics pipelines work based on multivariate classifiers, and as a result, more than one cell population can be responsible for the final predictions. Therefore, markers can have different relative importance in defining the multiple cell populations within the multivariate model. Quantifying the markers for each phenotype involved in the multivariate model can provide additional insight into the differences between closely related cell populations. For example, if two phenotypes $1^+2^-3^+4^-5^+$ and $1^+2^-3^+4^-6^+$ are identified as correlates of a disease, and if markers 5 and 6 (which are the only differences between them) are the least important markers for the former and latter phenotypes respectively, then these two phenotypes are likely to correspond to the same cell population (as far as the correlation with the disease is concerned). However, if markers 5 and 6 are the most important for the phenotypes, these can correspond to two biologically different cell populations.

To address these problems, we developed RchyOptimyx, a computational tool that uses dynamic programming and optimization techniques from graph theory to construct a cellular hierarchy, providing the best gating strategies to identify target populations to a desired level of purity or correlation with a clinical outcome, using the simplest possible combination of markers.

Materials and Methods

Our methodology builds on the flowType pipeline(27). flowType comprehensively identifies cell populations defined by all possible gating strategies (hierarchies) in the data set using a partitioning strategy (e.g., clustering algorithm like flowMeans (27)) and scores them by a statistical test (e.g., the log rank test for difference in survival distributions). Given the list of all cell populations and their scores, RchyOptimyx uses a dynamic programming approach to find the best cellular hierarchy within a reasonable time for interactive data analysis (e.g., less than 2 min for 30 color data), as well as a number of best suboptimal hierarchies, to enable mining of the space of best gating strategies and purities for a given target cell population.

Terms and Definitions

Let $\mathcal{M}$ be the set of m markers of interest (e.g., $\mathcal{M} = \{KI\text{–}67, CD28, CD45RO\}$), a single marker phenotype be a phenotype having only one marker (e.g., $CD28^+$), a phenotype P be a set of single marker phenotypes (e.g., $P = KI\text{–}67^+CD28^-$), and M (not to be mistaken with $\mathcal{M}$) be a phenotype of size m that involves all of the markers (e.g., M = KI–67^+CD28^-CD45RO$^-$). The power set of M, $\mathcal{P}$(M), is of size

[1] http://flowcap.flowsite.org/summit2011.html

2^m and contains every possible subset of M. The scoring function $S(.)$ assigns a score to each member of $\mathcal{P}(\mathrm{M})$, such that higher values are assigned to more important phenotypes (e.g., those with a stronger correlation with a clinical outcome).

Given an arbitrary M, the directed acyclic graph (DAG) G_M has $m + 1$ levels from 0 to m, each level i including every member of $\mathcal{P}(\mathrm{M})$ of size i. Node s is connected to node t with a directed edge (s,t) if and only if $|t| = |s| + 1$ and the two associated sets of s and t differ only in one single phenotype marker (i.e., t is an immediate parent of s). Let the weight of edge (s,t) be $-S(t)$ (so that paths with maximum score can be found by searching for paths with minimum total weight).

The node with 0 markers is the root (or source) node, and the node with the complete set of markers is the sink node. A path from source to sink is called a hierarchy path, or simply a hierarchy. An example of graph G_M for M = KI–67$^+$CD4$^-$CCR5$^+$CD127$^-$ is illustrated in Supporting Information Figure S3.

The graph G_M has $|\mathcal{P}(\mathrm{M})| = 2^m$ nodes, one node for each parent phenotype of the phenotype of interest. The number of edges is equal to the number of markers (m), times the number of edges that have the specified marker. Each marker appears in 2^{m-1} nodes, therefore the number of edges is $m \times 2^{m-1}$.

A scoring function is needed to find the best hierarchy. This function should give a higher rank to hierarchies that go through more important parent populations earlier (i.e., those that achieve a higher clinical significance with fewer markers). Because each node of the hierarchy is a phenotype, and each phenotype has a given score value $S(.)$, we use the total score function $T(.)$—the sum of all negated phenotype scores in the hierarchy—as the scoring function:

$$T(\mathcal{H}) = \sum_{(s,t)\in E_\mathcal{H}} W(s,t) = \sum_{(s,t)\in E_\mathcal{H}} -S(t) = \sum_{t\in V_\mathcal{H}\setminus v_0} -S(t) \quad (1)$$

where $\mathcal{H}$ is the given hierarchy, $E_\mathcal{H}$ is the set of edges of hierarchy $\mathcal{H}$, $V_\mathcal{H}$ is the set of vertices of same hierarchy, and v_0 is the first node in the hierarchy. Applying this function to G_M, the best hierarchy is the minimum weighted path in G_M. We note that, in principle, more complex functions can be used to compute the total score of a given hierarchy; for example, in applications in which phenotypes with fewer markers are more important than the other phenotypes, an exponential function can be used to increase the weight of the earlier phenotypes in the hierarchy.

Dynamic Programming to Identify The Best Hierarchy

For cell populations characterized by m markers, finding the best hierarchy by searching through all possible hierarchies would require time $O(m!)$, which is impractical for even moderately large m. To make this problem tractable using dynamic programming, we define best total score function $T^*(.)$, which computes the score of the best hierarchy leading to the given phenotype. $T^*(.)$ is defined recursively as follows:

$$T^*(P^k) = \begin{cases} -S(P^k) & \text{if } k = 1 \\ \min\{T^*(P^k \setminus P^k_i) - S(P^k) | i = 1,\ldots,k\} & \text{otherwise} \end{cases}, \quad (2)$$

where P^k is a cell population defined by k single marker phenotypes, and $P^k\backslash P^k_i$ is P^k with the ith single marker phenotype removed. For example, if P^3 = KI–67$^+$CD28$^-$CD45RO$^+$, then $P^3\backslash P^3_1 = CD28^- CD45RO^+$. In other words, there is an edge from $P^k\backslash P^k_i$ to P^k in G_M where, P^k is a subset of M. Also note that $-S(P^k)$ is the weight of the edge $(P^k, P^k\backslash P^k_i)$ in G_M. Using dynamic programming, we calculate the value of $T^*(.)$, iterating from level 0 to m on G_M. Calculating each node's score requires a number of constant time operations equal to the number of edges entering the node. Therefore, the total number of operations is proportional to total number of edges ($m \times 2^{m-1}$), and the overall time complexity of our programming procedure for determining $T^*(.)$ values for all phenotypes in the graph is $O(m \times 2^{m-1})$. An illustration of the dynamic programming space for three dimensional space, that is, having three markers, as well as two paths in that space is shown in Figure 1.

Search for Near-Optimal Hierarchies

The hierarchy selected by the dynamic programming algorithm is the best gating strategy for a given cell population. However, we would also like to identify alternate gating strategies with slightly worse scores. To find these near-optimal paths, we reformulate the problem as identification of a desired number of minimum weight paths: In G_M, the minimum weight path from source to sink is the best hierarchy (identical to the one generated by dynamic programming). To generate additional, suboptimal hierarchies, a list of the next minimum weight paths must also be generated. These paths can be identified using the method by Eppstein (33). As noted in the original article, elaborating the details of this algorithm is complicated and requires substantial background in algorithm design, which is well beyond the scope of this work. Briefly, this method uses the minimum spanning tree of G_M and computes a heap structure for each node; it then merges the heaps in an efficient way to construct a 4-heap data structure. Using this 4-heap and a given arbitrary number l (the number of desired paths), it generates l-minimum weight paths in time $O(e + v + l)$ for a DAG with e edges and v nodes [see Theorem 4 of (33) for details].

Hence, the time complexity of our algorithm can be calculated based on the number of edges and nodes using the time complexity of the l-minimum weight paths method:

$$\begin{aligned} O(e+v+l) &= O(m\times 2^{m-1} + 2^m + l) \\ &= O(m\times 2^{m-1} + 2\times 2^{m-1} + l) \\ &= O((m+2)\times 2^{m-1} + l). \end{aligned} \quad (3)$$

For example, the number of operations with our approach on a dataset with $m = 10$ markers would be $\approx 10^4$ compared to

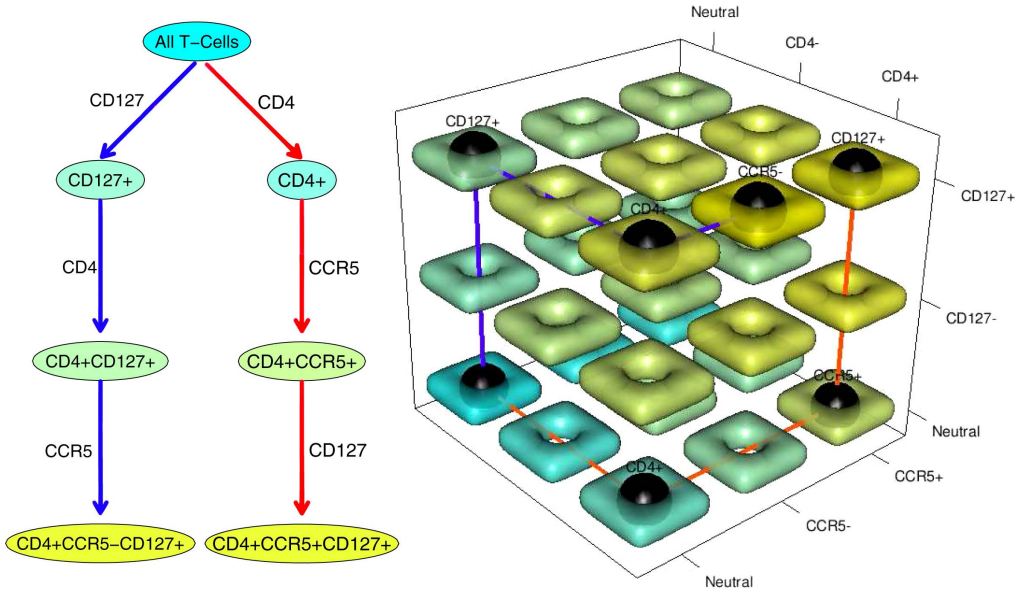

Figure 1. Dynamic programming algorithm for two cell populations defined by three markers. The best path for each of the cell population is shown in red and blue respectively. As an example, the red path ends at $CD4^{+}CCR5^{+}CD127^{+}$. Three markers are available to be added. First, CD4 is added (changes from does not matter to positive). Then two options will be available for the next step (CD127 and CCR5). After selection of CCR5, only one option will be left for the final step (CD127). Therefore for three markers, $\frac{3\cdot(3-1)}{2} = 6$ comparisons were required. Left: A hierarchy for the two paths. The label of an edge is the name of the single marker phenotype that is the difference between its head set (s) and its tail set (t). Right: the dynamic programming space for the three markers. Black spheres mark the nodes in the dynamic programing space used by the two paths. The colors of the nodes on the left match that of the square tori on the right and correspond to the relative score of each cell population. [Color figure can be viewed in the online issue which is available at wileyonlinelibrary.com]

$\approx 3 \times 10^6$ for the exhaustive search approach. Our method therefore takes ≈ 0.23 CPU seconds vs. ≈ 69 CPU seconds for exhaustive search, run under 64 bit Linux (version 3.3) on $2.93GHz$ Intel Xeon CPU with sufficient memory (proportional to 2^M). For a phenotype involving $m = 20$ markers, these numbers increase to ≈ 1.2 CPU seconds vs. $\approx 10^{11}$ CPU seconds [more than 4000 years), respectively. Even for a phenotype involving $m = 30$ markers measured by a CyTOF assay (mass spectrometry-flow cytometry hybrid device (25,34,35)], RchyOptimyx remains feasible, with a runtime of ≈ 102 CPU seconds, while the brute-force method would take $\approx 10^{22}$ CPU seconds. The final output of RchyOptimyx is the corresponding subgraph of G_M that includes all calculated paths (i.e., the optimized hierarchy, e.g., Supporting Information Fig. S4).

Datasets

We validated RchyOptimyx on two high-dimensional datasets, produced by mass and polychromatic flow cytometry.

Mass Cytometry Analysis of Bone Marrow Cells from Normal Donors

In this dataset, 31 parameters were measured for mononuclear cells from a healthy human bone marrow (see Ref. 25 for details). We used the results of three assays on samples subject to ex vivo stimulation by IL7 (measured by pSTAT5), BCR (measured by pBLNK), and LPS (measured by p-p38) as well as an unstimulated control. Thirteen surface markers were included in the analysis: CD3, CD45, CD45RA, CD19, CD11b, CD4, CD8, CD20, CD34, CD33, CD123, CD38, and CD90. Singlets were gated manually, as described in the original publication.

Polychromatic Flow Cytometry Analysis of HIV$^+$ Patients

This dataset consists of 13 color PFC assays of 466 HIV^+ subjects enrolled in the Infectious Disease Clinical Research Program's HIV Natural History Study. Basic demographic characteristics of this dataset are described elsewhere (36). Cryopreserved peripheral blood mononuclear cells stored within 18 months of the date of seroconversion were analyzed using PFC as described by Ganesan et al. (37). The cohort included 135 death/AIDS events, as defined by 1993 guidelines (38). The date of the last follow-up or initiation of highly active anti-retroviral therapy (HAART) was considered a censoring event. CD14 and V-amine dye were used to exclude monocytes and dead cells, respectively, CD3 was used to gate T-cells. Using the staining panel and flowType, we enumerated various subsets of naive and memory T-cells, defined by CD4, CD8, CD45RO, CD27, CD28, CD57, CCR5, CCR7, CD127, and KI-67. Using a log rank test with Bonferroni's multiple test correction, we scored each subset (cell population) in terms of its correlation with HIV progression (27).

Results

Designing a Panel to Detect a Population Expressing an Intracellular Marker using Surface Markers

In this use-case, our goal was to identify cell populations that are affected by different stimulations in the mass cytometry

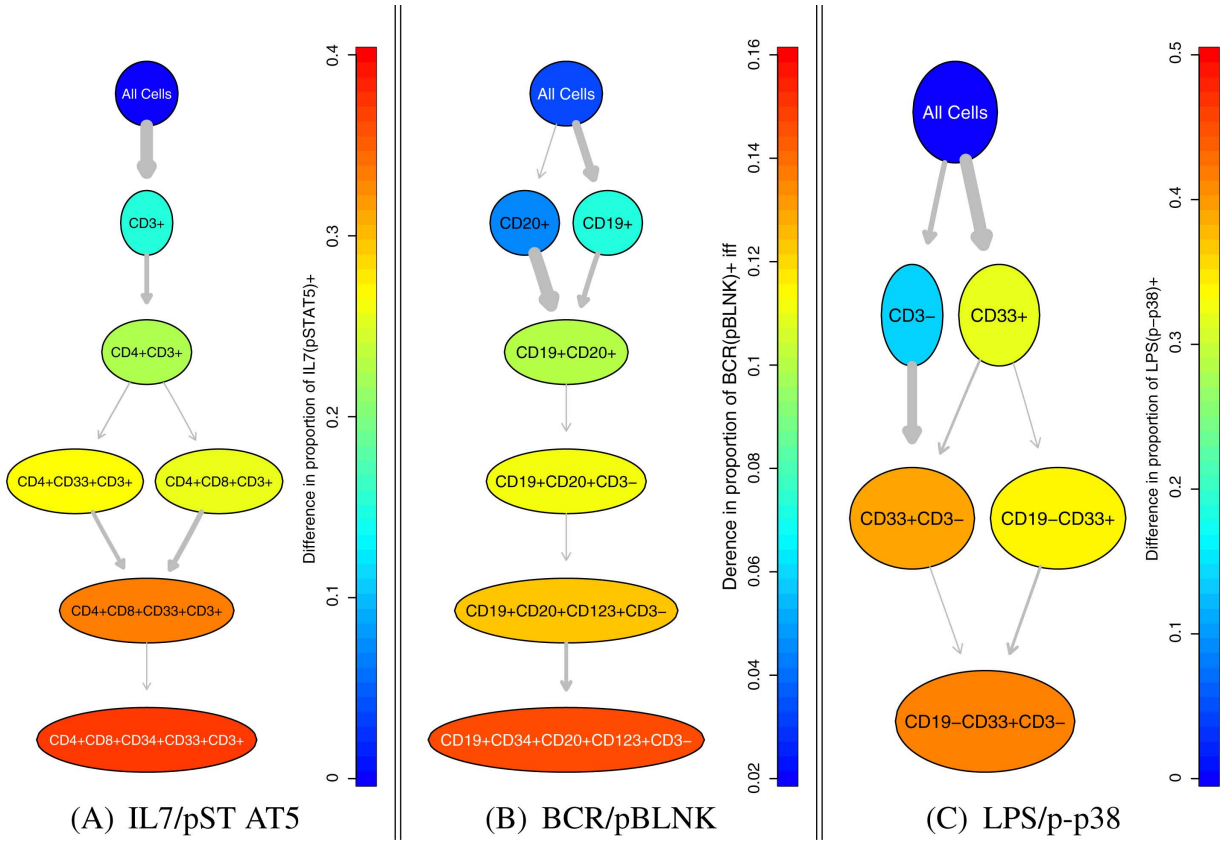

Figure 2. Three optimized hierarchies for identification of cell populations with maximum response to IL7, BCR, and LPS measured by pSTAT5, pBLNK, and p-p38, respectively. The color of the nodes and the thickness of the edges shows the proportion and change in proportion of cells expressing the intracellular marker of interest, respectively. [Color figure can be viewed in the online issue which is available at wileyonlinelibrary.com]

dataset. We used flowType to identify a list of populations that had a high overlap with either the $IL3^+$, BCR^+, or LPS^+ populations (determined manually, see Supporting Information Fig. S6). For each cell population, this value was calculated as the difference in its intersection with the $IL3^+$, BCR^+, or LPS^+ compartments between the stimulated and unstimulated sample. For example, for a given cell population CP, the overlap with $IL3^+$ was defined as:

$$\text{Overlap}^{IL3^+}(\text{CP}) = \left(\frac{\#\ IL3^+\text{cells in CP}}{\#\ \text{cells in CP}}\right)_{\text{stim}} - \left(\frac{\#\ IL3^+\text{cells in CP}}{\#\ \text{cells in CP}}\right)_{\text{unstim}} \quad (4)$$

The immunophenotypes with a high overlap, as identified by flowType, are listed in Supporting Information Tables S1–S3. These immunophenotypes were analyzed using RchyOptimyx (e.g., Supporting Information Fig. S1 for BCR) and then merged into a single graph, shown in Figure 2. This graph suggests that T-cells ($CD3^+$) followed by cytotoxic T-cells ($CD3^+CD4^+$) are the main parent populations that are affected by IL7 stimulation (panel A). As expected, BCR stimulation affected B-cells ($CD19^+CD20^+CD3^-$), and LPS stimulation increased the proportion of $CD19^-CD33^+CD3^-$ cells (Panels B and C, respectively). These results are generally consistent with those reported in the original study (Fig. 2 and panel C of Fig. 3 of Ref. 25).

Simplifying Gating Strategies

Here, we use RchyOptimyx to demonstrate an example of the use case of establishing a simpler combination of markers that can be used to identify a target population at a desired level of purity. For analysis of the PFC dataset, Ganesan et al. used a strict, but potentially redundant definition for naive T-cells, of $CD28^+CD45RO^-CD57^-CCR5^-CD27^+CCR7^+$, within the $CD3^+CD14^-$ compartment (37). The purity of a given parent cell population (CP) of this target was defined as its mean purity for the strictly-defined naive T-cells:

$$\text{Purity}(\text{CP}) = \frac{\sum_{\text{All Samples}} \frac{\#\ CD28^+CD45RO^-CD57^-CCR5^-CD27^+CCR7^+\ \text{cells}}{\#\ \text{cells in CP}}}{\#\ \text{Samples}} \quad (5)$$

Figure 3 shows the results of analysis with RchyOptimyx where a combination of only three markers ($CD45RO^-CCR5^-CCR7^+$) identified the strict naive T cell population to 95% purity (within the $CD3^+CD14^-$ compartment). The range of available purities, and determination of an appropriate cutoff is experiment dependent (e.g., on the range of available markers, biological

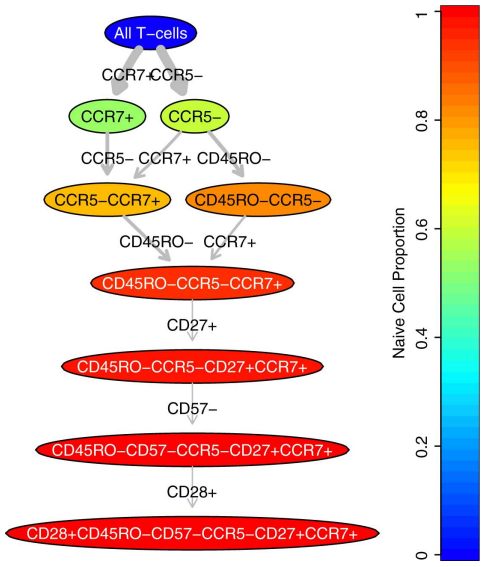

Figure 3. An optimized cellular hierarchy for identifying naive T-cells. The color of the nodes and the thickness of the edges shows the purity and change in purity of the original naive phenotype within the given cell population, respectively. [Color figure can be viewed in the online issue which is available at wileyonlinelibrary.com]

question being researched) and this result is only provided as an example of the utility.

Characterization of a Large Number of Immunophenotypes

Here, we use RchyOptimyx to demonstrate an example of the use-case of summarizing a large list of immunophenotypes of interest (as identified by a bioinformatics pipeline) into a single hierarchy using their most important common parent populations.

In a previous study of the PFC dataset, we identified 101 immunophenotypes (Supporting Information Table S4) in HIV^+ patients that had a statistically significant correlation with HIV's progression (27). The score of each population was calculated as $-\log_{10}(P)$ where P was the P-value of the logrank test before adjustment for multiple testing (higher values represent a stronger correlation with the clinical outcome). The 101 immunophenotypes were analyzed using RchyOptimyx and the resulting hierarchies were merged into a single graph (Fig. 4). This graph indicated three groups of immunophenotypes that were significantly correlated with HIV's outcome (left, center, and right branches). The left branch consisted of KI-$67^+CD4^-CCR5^+CD127^-$ T-cells. These cells were thought to be statistical significant mainly because they are long-lived ($CD127^-$) T-cells with high proliferation (KI-67^+). RchyOptimyx showed that the significance of this population is related to the KI-67^+CCR5^+ compartment and not $CD127^-$ (Fig. 4, the left branch) as the CD127 marker is not needed to achieve the approximately the same score. This is in agreement with the results of two recent studies (39,40). Theterminal node of the center branch consisted of seven markers ($CD45RO^-CD8^+CD57^+CCR5^-CD27^+CCR7^-CD127^-$). RchyOptimyx revealed that its most important parent population is $CD8^+CCR7^-CD127^-$, with a weaker correlation with the clinical outcome. Finally, the right branch ($CD28^-CD45RO^+CD4^-CD57^-$ $CD27^-CD127^-$) suggests several parent populations with minimal overlap and strong correlation with the clinical outcome (e.g., $CD28^-CD4^-CD57^-CD127^-$ and $CD45RO^+$ $CD4^-CD127^-$).

Discussion

Sequential analysis of the markers involved in manual or automated identification of cell populations is fundamental to our understanding of the characteristics of the cell population. In sequential gating, the order in which the gates have been applied does not affect the final results. However, ordering the gates by their relative importance has two use-cases: (1) identifying a cell population of interest, using the smallest possible panel of markers; (2) summarizing a long list of closely related (and perhaps overlapping) immunophenotypes by identifying their most important common parent populations. However, increasing the number of markers quickly renders this approach unfeasible (e.g., Supporting Information Fig. S5 for only six markers).

To address this challenge, we developed RchyOptimyx, a computational tool that automatically characterizes the complex findings of high-dimensional exploratory FCM studies. RchyOptimyx sorts all parent populations of an immunophenotype of interest into hierarchies, and selects those hierarchies that are better able to maintain the characteristics of the immunophenotype of interest (e.g., correlation with a clinical outcome). This reveals the best order in which markers can be excluded from an immunophenotype. RchyOptimyx uses dynamic programming and efficient tools from graph theory to make the problem tractable using the computing resources readily available in most laboratories.

As most cells can be described using more than one combination of markers, there usually are several alternative cellular hierarchies associated with every population. RchyOptimyx is able to find all these "paths" and merge them into a single hierarchy, starting from "all cells," or any arbitrary point in a hierarchy, and finishing at the terminal population of interest. This reveals the relationships between different gating strategies and how they differentiate, and also facilitates the reproduction of high-dimensional exploratory studies using low-color instruments. The ability to suggest multiple panels is particularly important when designing new panels, because the choice of markers depends on a large number of external parameters including, but not limited to, reagents available through vendors, potential spectral overlaps, the instruments available, and budget limitations.

Another important use-case for RchyOptimyx is in the interpretation of the findings of bioinformatics pipelines. Although these pipelines have recently been very successful in identifying cell populations correlated with clinical outcomes, their findings cannot be easily understood for two reasons: (1) they usually rely on high-dimensional clustering of the data and therefore cannot propose gating strategies for reproduction of their results; (2) their predictive power often relies on a large list of immunophenotypes. Some of these immunophenotypes

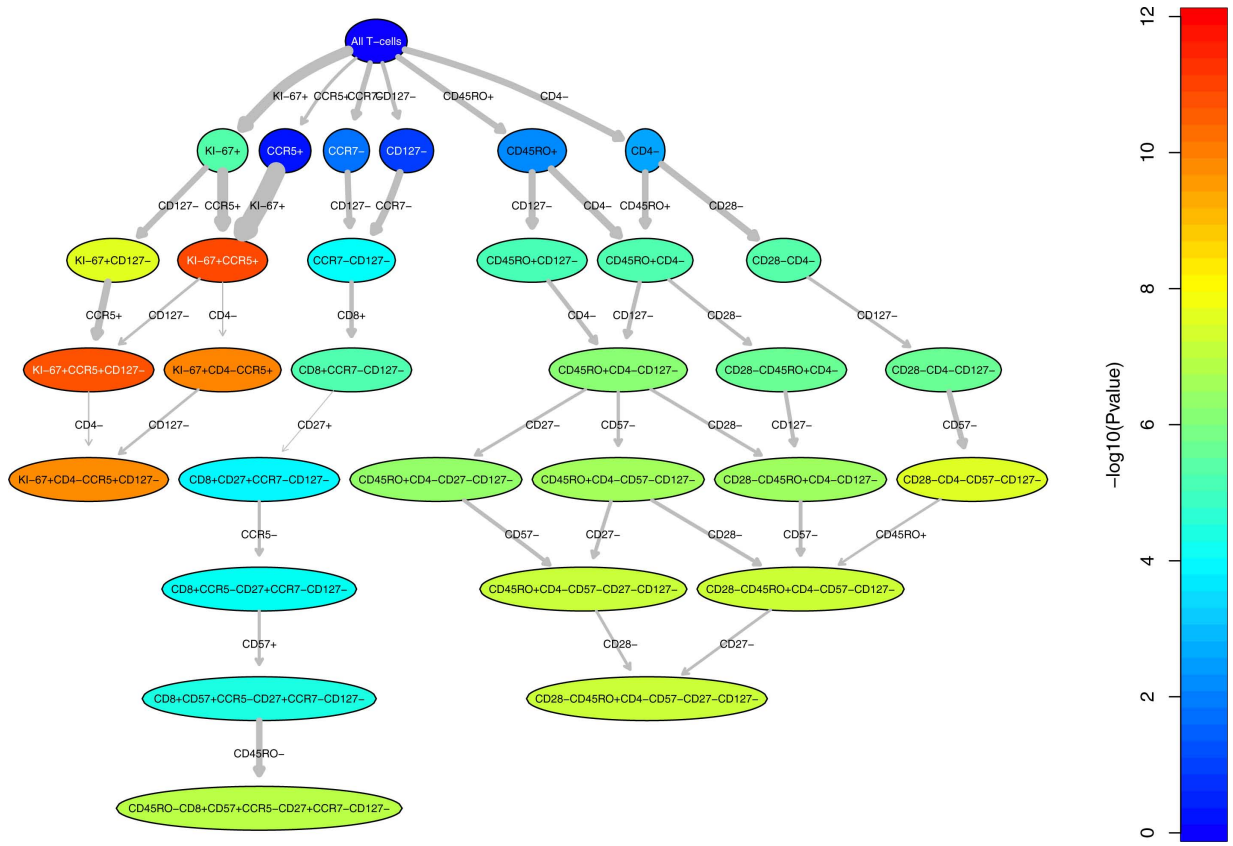

Figure 4. An optimized hierarchy for all three populations correlated with protection against HIV. The color of the nodes shows the significance of the correlation with the clinical outcome (*P*-value of the logrank test for the Cox proportional hazards model) and the width of each edge (arrow) shows the amount of change in this variable between the respective nodes. The positive and negative correlation of each immunophenotype with outcome can be seen from the arrow type leading to the node; however as all correlations are negative in this hierarchy, only one arrow type is shown. [Color figure can be viewed in the online issue which is available at wileyonlinelibrary.com]

are closely related (e.g., refer to close or overlapping cell populations) while others are not. RchyOptimyx addresses the first problem by suggesting optimized gating hierarchies for identification of these cell populations to a desired level of purity or correlation with clinical outcome. The latter problem is addressed by summarizing closely related immunophenotypes using their most important common parents.

In evaluating RchyOptimyx, we combined its functionality with the automated gating functionality provided by flowMeans and flowType. However, RchyOptimyx can be built upon the results of any cell population identification method, including manual analysis, provided all intermediate cell populations (i.e., each layer, removing one marker at a time) from the cell population of interest up to the desired start of the hierarchy are provided to the algorithm.

We evaluated RchyOptimyx for three use-cases, using a small but high-dimensional mass cytometry dataset and a clinical dataset of high-dimensional conventional FCM assays of 466 patients, previously analyzed by both manual and automated analysis. First, we constructed cellular hierarchies for identification of cells that were produced in response to different stimulations. This use-case represents the problem of designing panels of surface markers (primarily for sorting) for cells that can only be defined using their intra-cellular signature (possibly after proper stimulation). For example, plasmacytoid dendritic cell (PDC)s are known to express the toll-like receptor 9 (TLR9) in response to stimulation using CpG (41). A large number of surface candidates were recently proposed for PDCs (42–45). An interesting direction to extend this work would be to measure all these markers in a single panel, subject to CpG stimulation (using appropriate controls) to design a panel of surface markers for PDCs. In this case, TLR9 could be used as the external variable for optimization.

Second, we demonstrated that RchyOptimyx can be used to simplify existing gating strategies, using as an example the identification of naive T-cells previously defined using a complex panel of six markers to a 95% purity using only three. This proof-of-concept use-case is relevant when a subset of markers needs to be selected for reproduction of the results using fewer colors. For certain biological use–cases, purity of higher than 95% can be required. For such use-cases, a larger number of markers for exclusion of non-naive T-cells should be included in the panel.

Third, we showed that RchyOptimyx, together with a complex bioinformatics pipeline, can analyze a large high-dimensional clinical dataset, to reveal correlates of a clinical

outcome, hidden from previous manual and automated analysis of the same dataset. In addition, RchyOptimyx suggests the best gating strategies and marker panels for reproduction of these results in low-color settings. By identifying the best cellular hierarchies, RchyOptimyx allows the user to make an informed decision about the trade-off between the number of markers and the significance of the correlation with the clinical outcome. This feature is particularly important in hypothesis generating studies that need to be further validated using large clinical studies.

For the third example, it is important to note that the correct measure for the amount of correlation with a clinical outcome is an effect size (such as the root squared error of the estimated proportional hazard). However, such effect size does not provide any information about the significance of the correlation. As RchyOptimyx is intended to be a decision support tool, and in this case the decision is the degree to which a cell population can be generalized while maintaining the statistical significance of the correlation, we decided that the *P*-values of the log-rank tests were more appropriate for optimization of the hierarchies. To support this decision, we empirically investigated the differences between the *P*-values and effect sizes of the Cox proportional hazard models (Supporting Information Fig. S7) and concluded that these values are highly correlated (which is not surprising considering the large size of our cohort). It should be noted that as RchyOptimyx allows the user to choose which measure to provide, they can make this decision as appropriate for their specific data.

The concept of computationally extracting cellular hierarchies from FCM data has previously been introduced by the SPADE algorithm (25,26). SPADE generates a large number of multidimensional clusters and then connects them to each other using the distance between their mean/median fluorescence intensities. These are then manually annotated by biologists with domain knowledge. This makes SPADE useful for identification and visualization of a large number of clusters, particularly when expression of markers change gradually (e.g., cell-cycle analysis and some intracellular studies). However, the hierarchies generated by SPADE are logically and conceptually different from those generated by RchyOptimyx and have different use-cases. For example, the results of the mass cytometry dataset presented here are very close to results previously obtained from SPADE analysis. However, SPADE required manual annotation of the results by a human expert, using different plots demonstrating the expression of different surface markers and the intracellular marker of interest (Fig. 2 and panel C of Fig. 3 of Ref. 25). More complicated relationships that involve several markers cannot be easily identified by these manual annotations. In addition, SPADE is limited in that the relationships between cell populations is exclusively defined using the multidimensional distances between them. However, two cell populations that are close to each other in the multidimensional space can be far in terms of specific markers (which can be the most important ones). The cellular hierarchies generated by RchyOptimyx are based on parent-child relationships, guided by an external variable (cell populations that have common parents with similar patterns of correlation with a clinical outcome or intracellular response to stimulation are grouped together). This enables RchyOptimyx to automatically annotate a large number of cell populations identified by other methods (e.g., manual gating or SPADE) in terms of the importance of the markers involved and summarize them in a single hierarchy.

There are several directions in which this work can be extended. RchyOptimyx provides no information about the robustness of the hierarchies. Bootstrapping strategies could be used to produce confidence intervals for the tree structure and increase generalizability to previously unseen data (46). Also, our current implementation of RchyOptimyx assumes that every marker can be partitioned into a positive and negative population. Although the underlying theory does support additional (e.g., dim, bright, or low) populations, parts of the software package would need to be modified to accommodate these cases.

Availability

The RchyOptimyx R package (including source code, documentation, and examples) is freely available under an open source license (*Artistic 2.0*) and can be obtained from Bioconductor. The raw data and meta-data used in this study is publicly available through FlowRepository.org (under experiment ID *FR-FCM-ZZZK*) and through Cytobank.org (under experiment ID *6033*) for the PFC and CyTOF datasets, respectively.

Acknowledgments

The authors like to thank Greg Finak from the Fred Hutchinson Cancer Research Center for his comments on the RchyOptimyx Bioconductor package and Garry Nolan from the Stanford University for providing the mass cytometry dataset. Also, they like to thank the patients enrolled in the IDCRP Natural History Study without whom none of this work would have been possible. They like to thank the research coordinators and support staff who diligently work on the HIV Natural History Study, as well as the members of the Infectious Disease Clinical Research Program HIV Working Group.

Literature Cited

1. Maecker HT, McCoy JP, Nussenblatt R. Standardizing immunophenotyping for the Human Immunology Project. Nat Rev Immunol 2012;12:191–200.
2. Roederer M, Tárnok A. OMIPsOrchestrating multiplexity in polychromatic science. Cytometry Part A 2010;77A:811–812.
3. Mahnke, YD, Roederer, M. OMIP-001: Quality and phenotype of Ag-responsive human T-cells. Cytometry Part A 2010;77A:819–820.
4. Chattopadhyay P, Roederer M, Price D. OMIP-002: Phenotypic analysis of specific human CD8+ T-cells using peptide-MHC class I multimers for any of four epitopes. Cytometry Part A 2010;77A:821–822.
5. Wei C, Jung J, Sanz I. OMIP-003: Phenotypic analysis of human memory B cells. Cytometry Part A 2011;79A:894–896.
6. Biancotto A, Dagur P, Chris Fuchs J, Langweiler M, Philip McCoy J, Jr. OMIP-004: In-depth characterization of human T regulatory cells. Cytometry Part A 2011;81A:15–16.
7. Foulds K, Donaldson M, Roederer M. OMIP-005: Quality and phenotype of antigen-responsive rhesus macaque T cells. Cytometry Part A 2012;81A:360–361.
8. Murdoch D, Staats J, Weinhold K. OMIP-006: Phenotypic subset analysis of human T regulatory cells via polychromatic flow cytometry. Cytometry Part A 2012;81A:281–283.
9. Eller M, Currier J. OMIP-007: Phenotypic analysis of human natural killer cells. Cytometry Part A 2012;81A:447–449.
10. Zuleger C, Albertini M. OMIP-008: Measurement of Th1 and Th2 cytokine polyfunctionality of human T cells. Cytometry Part A 2012;81A:450–452.
11. Lamoreaux L, Koup R, and Roederer M. OMIP-009: Characterization of antigen-specific human T-cells. Cytometry Part A 2012;81A:362–363.

12. Preijers F, Huys E, Moshaver B. OMIP-010: A new 10-color monoclonal antibody panel for polychromatic immunophenotyping of small hematopoietic cell samples. Cytometry Part A 2012;81A:453–455.
13. Bendall S, Nolan G, Roederer M, Chattopadhyay P. A deep profiler's guide to cytometry. Trends Immunol 2012;33:323–332.
14. Perfetto S, Chattopadhyay P, Roederer M. Seventeen-colour flow cytometry: Unravelling the immune system. Nat Rev Immunol 2004;4:648–655.
15. Gattinoni L, Lugli E, Ji Y, Pos Z, Paulos C, Quigley M, Almeida J, Gostick E, Yu Z, Carpenito C, et al. A human memory t cell subset with stem cell-like properties. Nature Medicine 2011:p1290–1297.
16. Lo K, Brinkman R, Gottardo R. Automated gating of flow cytometry data via robust model-based clustering. Cytometry Part A 2008;73A:321–332.
17. Finak G, Bashashati A, Brinkman R, Gottardo R. Merging mixture components for cell population identification in flow cytometry. Adv Bioinf 2009:v09.
18. Pyne S, Hu X, Wang K, Rossin E, Lin T, Maier L, Baecher-Allan C, McLachlan G, Tamayo P, Hafler D, et al. Automated high-dimensional flow cytometric data analysis. Proc Natl Acad Sci, USA 2009;106:8519–8524.
19. Chan C, Feng F, Ottinger J, Foster D, West M, Kepler T. Statistical mixture modeling for cell subtype identification in flow cytometry. Cytometry Part A 2008;73A:693–701.
20. Naumann U, Luta G, Wand M. The curvHDR method for gating flow cytometry samples. BMC Bioinformatics 2010;11:11–44.
21. Zare H, Shooshtari P, Gupta A, Brinkman R. Data reduction for spectral clustering to analyze high throughput flow cytometry data. BMC Bioinformatics 2010;11:403–413.
22. Qian Y, Wei C, Eun-Hyung Lee F, Campbell J, Halliley J, Lee J, Cai J, Kong Y, Sadat E, Thomson E, et al. Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. Cytometry B Clin Cytom 2010;78(S1):S69–S82.
23. Sugár I, Sealfon S. Misty Mountain clustering: Application to fast unsupervised flow cytometry gating. BMC Bioinformatics 2010;11:502–508.
24. Aghaeepour N, Nikolic R, Hoos H, Brinkman R. Rapid cell population identification in flow cytometry data. Cytometry Part A 2011;79A:6–13.
25. Bendall S, Simonds E, Qiu P, Amir E, Krutzik P, Finck R, Bruggner R, Melamed R, Trejo A, Ornatsky O, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 2011;332:687–696.
26. Qiu P, Simonds E, Bendall S, Gibbs K, Jr. Bruggner R, Linderman M, Sachs K, Nolan G, Plevritis S. Extracting a cellular hierarchy from high-dimensional cytometry data with spade. Nat Biotechnol 2011;29:886–891.
27. Aghaeepour N, Chattopadhyay PK, Ganesan A, O'Neill K, Zare H, Jalali A, Hoos HH, Roederer M, Brinkman RR. Early Immunologic Correlates of HIV protection can be identified from computational analysis of complex multivariate T-cell flow cytometry assays. Bioinformatics 2012;28:1009–1016.
28. Zare H, Bashashati A, Kridel R, Aghaeepour N, Haffari G, Connors J, Gascoyne R, Gupta A, Brinkman R, Weng A. Automated analysis of multidimensional flow cytometry data improves diagnostic accuracy between mantle cell lymphoma and small lymphocytic lymphoma. Am J Clin Pathol 2012;137:75–85.
29. Costa E, Pedreira C, Barrena S, Lecrevisse Q, Flores J, Quijano S, Almeida J, del Carmen García-Macías M, Bottcher S, Van Dongen J, et al. Automated pattern-guided principal component analysis vs expert-based immunophenotypic classification of b-cell chronic lymphoproliferative disorders: a step forward in the standardization of clinical immunophenotyping. Leukemia 2010;24:1927–1933.
30. Roederer M, Nozzi J, Nason M. Spice: Exploration and analysis of post-cytometric complex multivariate datasets. Cytometry Part A 2011;79A:167–174.
31. Bashashati A, Johnson N, Khodabakhshi A, Whiteside M, Zare H, Scott D, Lo K, Gottardo R, Brinkman F, Connors J, et al. B cells with high side scatter parameter by flow cytometry correlate with inferior survival in diffuse large b-cell lymphoma. Am J Clin Pathol 2012;137:805–814.
32. Chan C, Lin L, Frelinger J, Hérbert V, Gagnon D, Landry C, Sékaly R, Enzor J, Staats J, Weinhold K, et al. Optimization of a highly standardized carboxyfluorescein succinimidyl ester flow cytometry panel and gating strategy design using discriminative information measure evaluation. Cytometry Part A 2010;77A:1126–1136.
33. Eppstein D. Finding the k shortest paths. SIAM J Comput 1998;28:652–673.
34. Ornatsky O, Bandura D, Baranov V, Nitz M, Winnik M, Tanner S. Highly multiparametric analysis by mass cytometry. J Immunol Methods 2010;361:1–20.
35. Chattopadhyay P, Roederer M. Cytometry: Today's technology and tomorrow's horizons. Methods 2012;57:251–258.
36. Weintrob A, Fieberg A, Agan B, Ganesan A, Crum-Cianflone N, Marconi V, Roediger M, Fraser S, Wegner S, Wortmann G. Increasing age at HIV seroconversion from 18 to 40 years is associated with favorable virologic and immunologic responses to HAART. J Acquir Immune Defic Syndr 2008;49:40–47.
37. Ganesan A, Chattopadhyay PK, Brodie TM, Qin J, Gu W, Mascola JR, Michael NL, Follmann DA, Roederer M, Decker C, et al. Immunologic and virologic events in early HIV infection predict subsequent rate of progression. J Infect Dis 2010;201:272–284.
38. Castro K, Ward J, Slutsker L, Buehler J, Jaffe H, Berkelman R, Curran J. Revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. MMWR Recomm Rep 1992;41:1–19.
39. Gordon S, Cervasi B, Odorizzi P, Silverman R, Aberra F, Ginsberg G, Estes J, Paiardini M, Frank I, Silvestri G. Disruption of intestinal CD4+ T cell homeostasis is a key marker of systemic CD4+ T cell activation in HIV-infected individuals. J Immunol 2010;185:5169–5179.
40. Jaspan H, Liebenberg L, Hanekom W, Burgers W, Coetzee D, Williamson A, Little F, Myer L, Coombs R, Sodora D, et al. Immune activation in the female genital tract during hiv infection predicts mucosal cd4 depletion and hiv shedding. J Infect Dis 2011;204:1550–1556.
41. Krug A, Towarowski A, Britsch S, Rothenfusser S, Hornung V, Bals R, Giese T, Engelmann H, Endres S, Krieg A, et al. Toll-like receptor expression reveals CpG DNA as a unique microbial stimulus for plasmacytoid dendritic cells which synergizes with CD40 ligand to induce high amounts of IL-12. European Journal of Immunology 2001;31:3026–3037.
42. Marafioti T, Paterson J, Ballabio E, Reichard K, Tedoldi S, Hollowood K, Dictor M, Hansmann M, Pileri S, Dyer M, et al. Novel markers of normal and neoplastic human plasmacytoid dendritic cells. Blood 2008;111:3778–3792.
43. Swiecki M, Colonna M. Unraveling the functions of plasmacytoid dendritic cells during viral infections, autoimmunity, and tolerance. Immunol Rev 2010;234:142–162.
44. Schuster P, Donhauser N, Pritschet K, Ries M, Haupt S, Kittan N, Korn K, Schmidt B. Co-ordinated regulation of plasmacytoid dendritic cell surface receptors upon stimulation with herpes simplex virus type 1. Immunology 2010;129:234–247.
45. Cao W. Molecular characterization of human plasmacytoid dendritic cells. J Clin Immunol 2009;29:257–264.
46. Suzuki R, Shimodaira H. Pvclust: an r package for assessing the uncertainty in hierarchical clustering. Bioinformatics 2006;22:1540–1542.