Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE

Peng Qiu[1,2], Erin F Simonds[3], Sean C Bendall[3], Kenneth D Gibbs Jr[3], Robert V Bruggner[3], Michael D Linderman[4], Karen Sachs[3], Garry P Nolan[3] & Sylvia K Plevritis[1]

The ability to analyze multiple single-cell parameters is critical for understanding cellular heterogeneity. Despite recent advances in measurement technology, methods for analyzing high-dimensional single-cell data are often subjective, labor intensive and require prior knowledge of the biological system. To objectively uncover cellular heterogeneity from single-cell measurements, we present a versatile computational approach, spanning-tree progression analysis of density-normalized events (SPADE). We applied SPADE to flow cytometry data of mouse bone marrow and to mass cytometry data of human bone marrow. In both cases, SPADE organized cells in a hierarchy of related phenotypes that partially recapitulated well-described patterns of hematopoiesis. We demonstrate that SPADE is robust to measurement noise and to the choice of cellular markers. SPADE facilitates the analysis of cellular heterogeneity, the identification of cell types and comparison of functional markers in response to perturbations.

Measurement of multiple parameters of single cells by flow and mass cytometry has diverse uses in understanding cellular differentiation and intracellular signaling cascades, clinical immunophenotyping, identifying rare stem cell populations and drug targeting using intracellular markers, such as phosphorylated proteins. Modern flow cytometers typically provide simultaneous single-cell measurements of up to 12 fluorescent parameters in routine cases, and analysis of up to 17 protein parameters has been reported[1]. Recently, the first commercially available next-generation mass cytometry platform (CyTOF) has become available and allows routine measurement of 30 or more single-cell parameters[2].

Despite the technological advances in acquiring an increasing number of parameters per single cell, methods for analyzing multidimensional single-cell data remain inadequate. Traditional methods are often subjective, labor intensive and require expert knowledge of the underlying cellular phenotypes. One common but cumbersome step is the selection of subsets of cells in a process called "gating"[3]. A gate is a region, defined in a biaxial plot of two measurements, which is used to select cells with a desired phenotype for downstream analysis. Gates are either manually drawn using software such as FlowJo and FlowCore[4], or automatically defined by clustering algorithms[5–10]. Manual gating is highly subjective and depends on the investigator's knowledge and interpretation of the experiment. Automatic gating algorithms cluster cells by optimizing the objective that cells in the same cluster be more similar to each other than cells from other clusters. Because these algorithms strive to define maximally different clusters, they often miss the underlying continuity of phenotypes (progression) that is inherent in cellular differentiation[11]. In addition, optimization objectives of most automatic gating algorithms are predisposed to capture the most abundant cell populations, whereas rare cell types, such as stem cells, are either excluded as outliers or absorbed by larger clusters. Some algorithms, such as a recent approach for automated gating termed SamSPECTRAL, have begun to include mechanisms for rare cell type identification[12].

Traditional cytometry data analysis methods also often cannot effectively accommodate and visualize the increasing numbers of measurements per single cell. For instance, to fully visualize an m-dimensional flow data set, $m(m-1)/2$ biaxial plots are needed, where each biaxial plot displays the correlation of only two measurements at a time. It is difficult to identify the correlations in high-dimensional data ($m \geq 3$) from a series of biaxial plots. One recent approach that partly addresses the scalability issue is the probability state model, implemented in the Gemstone software package. That approach rearranges cells into a nonbranching linear order, according to an investigator's knowledge or expectation of how known markers fluctuate along a progression underlying the measured cell population[13]. Because cells are ordered in a nonbranching fashion, a new model must be constructed for each mutually exclusive cell type (that is, T cells, B cells).

We developed the SPADE approach to extract a hierarchy from high-dimensional cytometry data in an unsupervised manner. SPADE is complementary to existing approaches for analyzing cytometry data by enabling multiple cell types to be visualized in a branched tree structure without requiring the user to define a known cellular ordering. Through a two-dimensional visualization, SPADE shows how measured protein markers behave across different cell types in the data; this empowers investigators to identify known cell types and to find unexpected ones. Recently, we reported the use of SPADE for immunophenotyping without providing a detailed description and

[1]Department of Radiology, Stanford University, Stanford, California, USA. [2]Department of Bioinformatics and Computational Biology, University of Texas, M.D. Anderson Cancer Center, Houston, Texas, USA. [3]Department of Microbiology and Immunology, Stanford University, Stanford, California, USA. [4]Computer Systems Laboratory, Stanford University, Stanford, California, USA. Correspondence should be addressed to P.Q. (pqiu@mdanderson.org).

Received 10 January; accepted 31 August; published online 2 October 2011; doi:10.1038/nbt.1991

analysis of the algorithm[14]. Here we detail the algorithm by which SPADE operates, perform comparison with traditional gating methods, evaluate its robustness, highlight its applications in identifying cell types and intracellular signal activations, and provide the source code of SPADE.

RESULTS

Outline of SPADE as applied to a simulated data set

To demonstrate the SPADE algorithm, we analyzed a simulated two-parameter flow cytometry data set (**Supplementary Data Set 1**). In the simulated cell population, the underlying cellular hierarchy originated from a rare 'root' cell type and differentiated into three distinct abundant cell types (**Fig. 1**). In a traditional gating analysis of the data set, four gates would be manually drawn, corresponding to the four distinct subpopulations (**Fig. 1**, i). Alternatively, SPADE views the data as a high-dimensional point cloud of cells, and uses topological methods to reveal the geometry of the cloud.

SPADE contains four computational modules. First, SPADE performs density-dependent down-sampling to equalize the density in different parts of the cloud and achieve equal representation of rare and abundant cell types (**Fig. 1**, ii). Second, SPADE performs agglomerative clustering to partition the down-sampled cloud into clusters of cells with similar phenotypes, that is, cells displaying similar intensities of the two markers (**Fig. 1**, iii). Because the down-sampling step makes the abundant and rare cell types relatively equally represented, the rare root cells are allowed to form their own clusters and not be outnumbered by abundant cell populations during clustering. Third, SPADE extracts and summarizes the geometry of the cloud by constructing a minimum spanning tree, the tree that connects all clusters with minimum total edge length (**Fig. 1**, iv). Finally, SPADE maps each cell in the original data set to the cluster in the tree to which it is most similar, a process called 'up-sampling'. As a result, properties of the cells in each cluster can be summarized and displayed on the tree. For instance, each node of the tree can be colored according to the median intensity of one of the simulated markers of the cells in that cluster, which allows visualization of the behavior of that marker across the entire heterogeneous cell population (**Fig. 1**, v). The two colored trees contain four branches with distinct phenotypes, shown by the manually drawn gray boundaries, which correspond to the four simulated cell types. In addition, the gradual change of marker intensity values along each lineage is evident.

When visualizing the SPADE tree, an important issue is to determine how the nodes and connections between them are to be arranged in a two-dimensional image. To aid in comprehension of the simulated two-parameter example, we defined the position of each node as the median intensities of the two markers of cells in that node. As a result, the raw data (**Fig. 1**, i) and the SPADE tree (**Fig. 1**, iv) are visually similar. For data with higher dimensions, SPADE uses a modified Fruchterman-Reingold algorithm to automatically compute the layout[15]. Detailed descriptions of the visualization algorithm and the four modules of SPADE are provided in Online Methods.

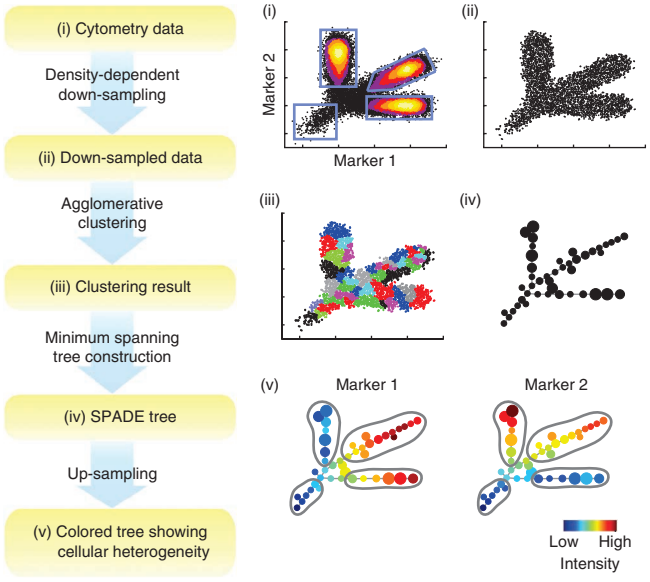

Figure 1 Flowchart of SPADE and SPADE analysis of a simulated data set. (i) A simulated two-parameter flow cytometry data set, with one rare population and three abundant populations. (ii) Result of density-dependent down-sampling of the original data. (iii) Agglomerative clustering result of the down-sampled cells. Adjacent clusters are drawn in alternating colors. (iv) Minimum spanning tree that connects the cell clusters. (v) Colored SPADE trees. Nodes are colored by the median intensities of protein markers of cells in each node, allowing visualization of the behaviors of the two markers across the entire heterogeneous cell population.

Analysis of mouse hematopoiesis using flow cytometry data

To validate the ability of SPADE to reconstruct a known branched cellular hierarchy, we used it to analyze a flow cytometry data set from a mouse bone marrow sample (**Supplementary Data Set 2**). Hematopoiesis in mice is well-described (**Fig. 2a**)[16,17], with multipotent self-renewing stem and progenitor cells giving rise to all of the terminally differentiated cell types. Mature myeloid cells are characterized by expression of the surface antigen CD11b, whereas lymphoid cells are negative for this marker. Within the lymphoid population, B cells express B220 but not TCRβ, and conversely the majority of T cells express TCRβ but not B220. Finally, mature TCRβ-expressing T-cells are characterized by mutually exclusive expression of CD4 or CD8.

When applied to the mouse bone marrow data set, SPADE required three user-specified input parameters. We chose the outlier and target densities empirically to be the 1st and 5th percentile of local densities of all the cells, and the desired number of clusters to be 50 (Online Methods). From these data and input parameters, SPADE automatically generated a tree diagram without annotations (**Fig. 2b**).

To interpret and annotate the SPADE tree, we created several versions of it, colored according to the median intensity of each measured marker. We used the colored trees to manually identify the type of cells represented by different parts of the tree. For example, in the tree colored according to c-kit intensity, the upper branch in the middle showed a clear pattern of high intensity. Therefore, this branch was annotated as c-kit$^+$ (**Fig. 2c**). Similarly, according to the SPADE tree colored by CD11b, the left branch showed high intensity. Based on the investigator's familiarity with this immunological system, this branch was labeled 'myeloids' (**Fig. 2d**). The remaining annotations labeled with B cells, T cells, dendritic cells, CD4$^+$ and CD8$^+$ were derived using similar logic (**Fig. 2e**). Notably, defining the boundaries of the annotations did not make use of gating or prior knowledge. From these annotations, we observed that different branches corresponded to different cell phenotypes. The interconnectivity among these phenotypes is consistent with known biology of mouse hematopoiesis.

To validate the annotations of the SPADE tree, we compared them with the result of expert-based traditional gating analysis, in which subpopulations of cells were identified by a series of gates manually drawn on biaxial plots (**Fig. 2f**). The manual gating analysis was performed in a blinded fashion before the SPADE analysis. For each gated

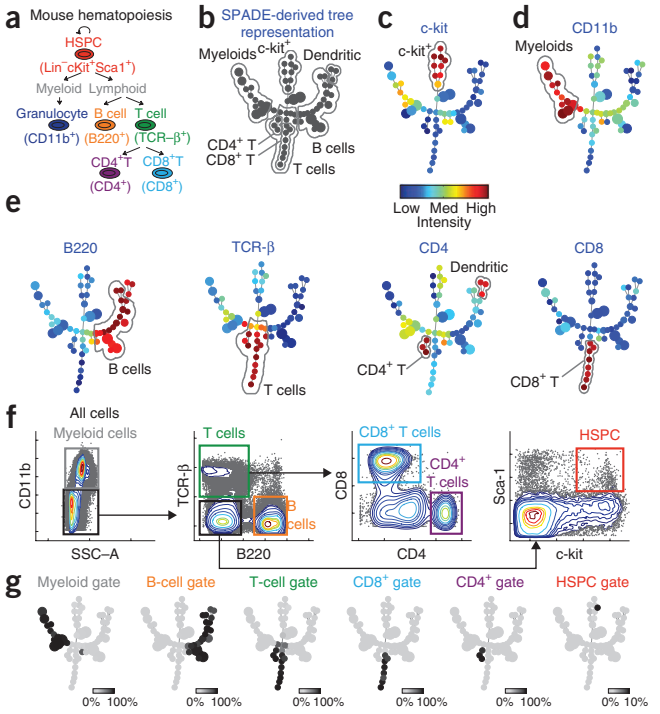

Figure 2 SPADE applied to mouse bone marrow flow cytometry data. (**a**) Known hematopoietic hierarchy in mouse bone marrow. (**b**) SPADE tree derived from the mouse bone marrow data. (**c**–**e**) Trees colored by the median intensity of one individual marker. (**f**) Traditional gating analysis on the mouse bone marrow data. (**g**) For each gated population, one SPADE tree was drawn, where each node was colored according to the percentage of gated cells in that node. Thus, the darker regions of each tree represent which part of the tree is populated by the cells in the corresponding gate. This comparison shows the concordance between SPADE and gating results.

population, we colored the tree by the percentages of the manually gated cells in each node, showing which part of the tree is populated by the cells in that gate (**Fig. 2g**). It can be observed that each gated population occupied one branch of the tree. Overall, the SPADE result was consistent with traditional gating analysis in identifying biologically relevant populations.

Notably, manual gating did not identify the dendritic cells because gating is a subjective approach that relies on our prior knowledge and we did not plan to find dendritic cells. Only after examining the SPADE results did we realize that manual gating could have been used to define a TCRβ$^-$ B220$^+$ CD4$^+$ dendritic cell population (**Supplementary Section 1**). In contrast, SPADE analysis readily identified the dendritic cell population as three nodes on the distal end of the B220$^+$ branch.

To quantify the difference between the two approaches, we computed the number of cells shared by all possible pairs of gates in the gating analysis and annotated regions in the SPADE tree (**Table 1**). Large values in the shaded entries indicate the consistency between gating and SPADE, whereas the differences are shown by the remaining entries. Cells in the B-cell gate were identified as B cells and dendritic cells in the SPADE tree (**Table 1**, column 1), consistent with **Figure 2**. The majority of cells in the myeloid gate were annotated by SPADE as myeloids, with a small fraction of B cells, T cells and c-kit$^+$ cells (**Table 1**, column 5). On the contrary, few cells in the myeloid region of SPADE were regarded as other cell types by gating (**Table 1**, row 6). In mouse bone marrow, c-kit is a marker for immature cell types, and the hematopoietic stem and progenitor cells (HSPCs) are a subset of c-kit$^+$ cells. The majority of cells in the manual HSPC gate belonged to the c-kit$^+$ branch (**Table 1**, column 6) and were found to be localized to one node in that branch of the SPADE tree (**Fig. 2g**).

We performed two analyses to evaluate the robustness of SPADE. First, to evaluate how marker selection affects the SPADE tree, we applied SPADE to reconstruct the mouse bone marrow hierarchy based on data sets consisting of subsets of the measured markers. We initiated SPADE to analyze only one marker, incrementally added more markers and evaluated changes in the resulting SPADE tree. These analyses demonstrated that the SPADE tree is only altered by markers that provide a sufficient amount of new information to change the shape of the cloud but not by markers that are highly correlated to the ones already included to build the tree (**Supplementary Section 2**). Second, to further evaluate the robustness of SPADE, we simulated new data by adding noise to the mouse bone marrow data, which already contains experimental noise. Our simulation suggests that SPADE can tolerate a small amount of additional noise. When the s.d. of the added noise was 5% of that of the data, even though parts of the SPADE tree inevitably varied, the overall topology and general interpretation were not affected (**Supplementary Section 3**).

Analysis of human hematopoiesis using mass cytometry data

Next-generation mass cytometry technology currently provides simultaneous measurement of 31 or more markers per cell. Such a capacity allows enough surface markers to delineate nearly all cell types in human hematopoiesis, as well as additional functional markers to study cellular response to perturbations. Previously we generated a mass cytometry data set of human bone marrow[14]. Single-cell measurements of 30 individual experiments were obtained (**Fig. 3a**). One unstimulated aliquot of the human bone marrow sample was measured with an immunophenotyping panel of 31 cell surface antibodies. In addition, we measured 5 unstimulated samples and 24 samples under different perturbations, using a functional staining panel of

Table 1 Comparison of manual gating and SPADE

Annotated SPADE branches	Gates in the gating analysis					
	B cell (150,314)	T cell (14,699)	CD4$^+$ (2,808)	CD8$^+$ (6,055)	Myeloid (209,079)	HSPC (418)
B cell (152,685)	146,017 (97.1%)	88 (0.6%)	0	0	2,246 (1.1%)	16 (3.8%)
Dendritic (3,996)	3,562 (2.4%)	79 (0.5%)	77 (2.7%)	0	83 (<0.1%)	0
T cell (17,538)	364 (0.2%)	12,377 (84.2%)	2,729 (97.2%)	6,037 (99.7%)	3,033 (1.5%)	0
CD4$^+$ (2,931)	0	2,858 (19.4%)	2,713 (96.6%)	0	27 (<0.1%)	0
CD8$^+$ (6,301)	0	6,174 (42%)	0	5,843 (96.5%)	32 (<0.1%)	0
Myeloid (202,180)	5 (<0.1%)	75 (0.5%)	0	0	199,048 (95.2%)	0
c-kit$^+$ (15,681)	8 (<0.1%)	815 (5.5%)	2 (0.1%)	9 (<0.1%)	1,159 (0.6%)	401 (95.9%)

The total number of cells in each gate and each annotated SPADE region are provided. Each entry in the table gives the number of cells shared by a particular gate and a particular SPADE region. Percentages are defined as this number divided by the total number of cells in the corresponding gate, thereby representing the percent of cells in a gate that are assigned to each SPADE region. Large values in shaded entries indicate the consistency between manual gating and SPADE.

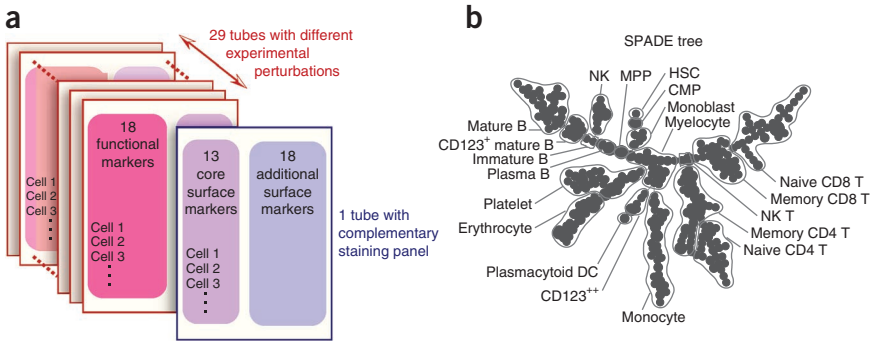

Figure 3 SPADE applied to human bone marrow data of 30 experiments with two overlapping staining panels and multiple experimental conditions. (**a**) Experiment and staining panel design. (**b**) SPADE tree derived from this data set. The SPADE tree was annotated according to its colored versions based on the 13 core surface markers. CMP, common myeloid progenitor; MPP, multipotent progenitor.

13 core surface markers (CD: 3, 4, 8, 11b, 19, 20, 33, 34, 38, 45, 45RA, 90, 123, from the 31-marker panel) and 18 intracellular targets that reflect intracellular signaling states. The 13 overlapping markers provide the opportunity to integrate the data of the two staining panels into a 49-dimensional data set. Here we detail the use of SPADE to integrate these two staining panels, and the use of all 49 dimensions to identify cell types and compare multiple perturbation conditions.

We first used SPADE to perform density-dependent down-sampling for each individual sample separately. To integrate the two staining panels, we applied the clustering step to the subset of the down-sampled data comprising the 13 overlapping core surface markers measured across down-sampled cells in the six unstimulated samples (Online Methods). The number of clusters was set to be 300, larger than that of the previous mouse bone marrow analysis, because the increased number of markers could capture more cell types and branch points. SPADE generated a tree (**Fig. 3b**), which we manually annotated by coloring the tree using each of the 13 core surface markers (**Supplementary Section 4**). The layout of the SPADE tree appears different from that reported previously[14] because the previous layout was manually organized to resemble the classic immunology diagram of hematopoietic developmental hierarchy[16], whereas the layout here was automatically generated (Online Methods).

Many classically defined immune cell subsets were immediately visible in the SPADE tree. Multiple nodes captured the abundant cell types, including B cells ($CD19^+$), T cells ($CD3^+$) and monocytes ($CD33^+$). In contrast, rare cell types, such as hematopoietic stem cells (HSC), only occupied a single node with high CD34 expression. The pattern of interconnectivity between these different cell types partially recapitulated established biology, as exemplified by the central positioning of the progenitor cell types, and the co-localization of multiple related T and B cell types. These results demonstrate the utility of SPADE to reduce a high-dimensional data set to an intuitive tree diagram that reflects the relatedness of biological subsets.

One particular group of nodes (**Fig. 4**) exhibited a consistent $CD38^+$ $CD45RA^+$ phenotype (**Supplementary Section 5**), but the identity of this cell type was not clear based on any of the 13 core surface markers from which the SPADE tree was built. As the SPADE tree was built using cells measured by both staining panels, we were able to use the 18 non-core surface markers in the immunophenotyping panel to color the SPADE tree. The unidentified nodes were found to be positive for CD7 and CD16 (**Fig. 4**), markers associated with natural killer (NK) cells. SPADE was able to cluster the NK cells without using NK-specific markers because the NK cells express a unique combination of the core surface markers $CD45^+$ $CD45RA^+$ $CD38^+$ $CD19^-$, which distinguishes them from other cell types. These results show that SPADE can identify a biologically relevant cell type from high-dimensional cytometry data, without using markers considered to be standard immunophenotypic indicators of that cell type.

We next discuss how the SPADE tree can be used to display the dynamics of intracellular markers under different perturbations. Integration of the two staining panels allowed the 18 function markers to be used to color the SPADE tree. For any combination of one intracellular marker and one perturbation, SPADE colored the tree according to the ratio between the median intensities of the marker in the stimulated and unstimulated (basal) conditions, showing the changes of the marker in response to the stimulation (**Fig. 5**). The activities of many functional markers supported the annotations derived from the surface markers. We analyzed all 432 SPADE trees colored by the measurements of 18 function markers across 24 different perturbation conditions. From those colored trees, we derived a distribution of s.d. of functional marker activities within the annotated boundaries. When we randomly permuted the tree nodes, we observed that the s.d. of functional markers' activities within the annotated boundaries was significantly smaller than random (two-sample student t-test $P < 10^{-25}$, **Supplementary Section 6**), thus verifying the relevance of the boundaries defined in **Figure 3b** to functional signaling responses in the cell.

Based on the SPADE trees colored by the activities of the functional markers, we observed multiple well-established signaling functionalities that were restricted to nodes with the expected manually annotated cell phenotypes. For example, TNF induction of phosphorylated MAPKAPK2 was observed in myeloid and NK cell types (**Fig. 5a**)[18]; the LPS-induced degradation of total IκBα, an indicator of NF-κB pathway activation, was restricted to cells of the monocytoid lineage, which uniquely express the receptor for LPS (**Fig. 5b**)[19]. We also observed evidence for two unreported hypotheses. First, the induction of phosphorylated STAT5 after stimulation with thrombopoietin (TPO) was expected in HSCs and earlier myeloid progenitors but not necessarily in the $CD123^{++}$ population (**Fig. 5c**). We inspected the raw data and confirmed the presence of a rare but well-defined $CD3^-$ $CD45RA^-$ $CD33^{mid}$ $CD38^+$ $CD123^{++}$ population that responded to TPO through phosphorylation of STAT5 (**Supplementary Section 7**). Although this immunophenotype does not match any reported

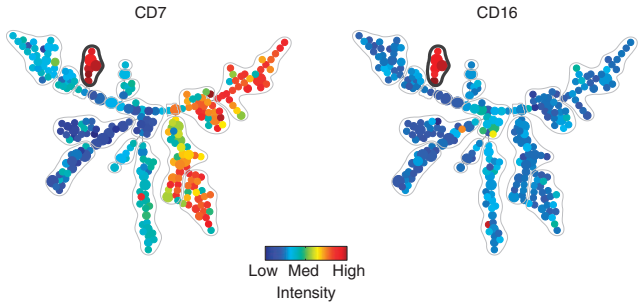

Figure 4 SPADE tree colored by two NK-specific markers CD7 and CD16, which were not used to derive the SPADE tree. The color patterns indicate that the nodes contained within the dark black boundary are NK cells.

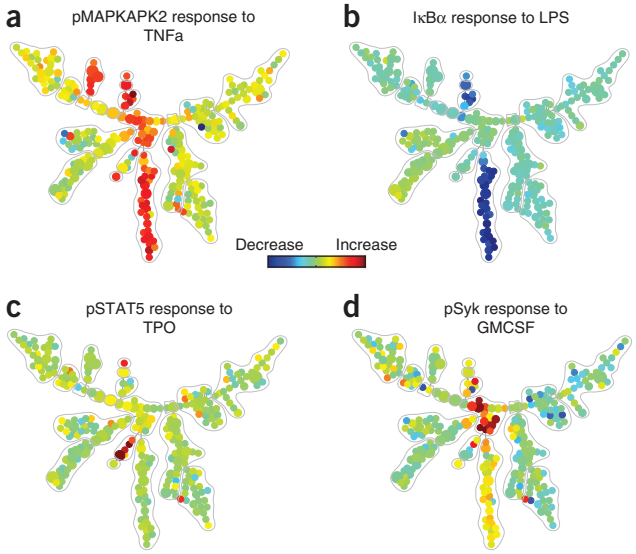

Figure 5 SPADE trees that describe the cell type–dependent behavior of functional markers in response to perturbations. (**a**) After stimulation with TNF, phosphorylated MAPKAPK2 was observed in myeloid and NK cell types, but not in other cell types. (**b**) After stimulation with LPS, degradation of total IκBα was restricted to the monocytoid lineage. (**c**) TPO-induced phosphorylated STAT5 was observed in HSCs and CD123^{++}, but not in other cell types. (**d**) GM-CSF–induced phosphorylation of pSyk was observed only in myelocytes.

immunological population based on the markers at hand, it may be a subset of dendritic cell progenitors, which has been previously described to exhibit enhanced *in vivo* expansion and maturation into plasmacytoid dendritic cells when TPO is added to the traditional Flt3-containing growth media[20]. Second, we observed GM-CSF-induced phosphorylation of pSyk in myelocytes (**Fig. 5d**). Similar signaling biology has been reported in neutrophils, which are the terminally differentiated progeny of myelocytes[19], but never directly reported in the bone marrow. This analysis demonstrates how SPADE can be used to map intracellular signal activation of functional markers across the landscape of human hematopoietic development.

DISCUSSION

SPADE enables the exploration of high-dimensional cytometry data in an objective manner that is scalable with increasing numbers of cellular parameters. More importantly, SPADE helps investigators infer likely cellular progressions and hierarchies. This can facilitate new biological discoveries, including the identification of unexpected signaling behaviors or the identification of rare cell types. We applied SPADE to a mouse bone marrow flow cytometry data set and a human bone marrow mass cytometry data set. In both data sets, SPADE was able to recover a hierarchy that illustrated known biology. In addition, we demonstrated that SPADE could be used to identify functionally distinct cell types and to study the activities of functional markers in response to perturbations.

The SPADE algorithm consists of four components (**Fig. 1**): density-dependent down-sampling, agglomerative clustering, linking clusters with a minimum spanning tree and up-sampling to restore all cells in the final result. This modular process allows more efficient sub-algorithms to replace the current components. In this sense, SPADE can be viewed as a framework for cytometric data analysis and visualization that has the capacity to be refined and adapted for new uses.

Algorithmically, SPADE is complementary to, and offers certain advantages over, traditional methods for analyzing cytometric data. First, SPADE does not require the user to impose a predefined hierarchical ordering of the cells using prior knowledge. Second, SPADE is suited for identifying rare cell types as it uses a density-dependent down-sampling scheme, which prevents the abundant cell types from dominating the statistics of the subsequent analysis. Finally, SPADE produces an easily visualized branching tree structure that in part recapitulates the branched cellular hierarchy that links related cell types. The resulting tree structure can be colored to display how surface and functional markers behave across the entire heterogeneous cell population.

The utility of SPADE is perhaps most limited by the choice of markers that are measured in the experiment and the subset of those that are used for building the SPADE tree. For instance, if the tree structure is built with a marker set that is not related to cellular progression, one might not expect to recover the known lineage relationships. In prior work on gene expression data analysis[21], we presented a potential approach for computationally selecting meaningful markers. Using a concept termed 'progression similarity', we identified subsets of genes that are concordant with a common hierarchical structure. As more markers can be measured on individual cells, this concept can be extended to cytometric data, as a means to select protein markers that support a common cellular hierarchy. In this manner, the utility of SPADE has the potential to increase as the number of markers per single cell increases. SPADE is intended to automatically produce intuitive representations of high-dimensional single-cell data that serve as exploratory tools for analysis.

In summary, SPADE is a versatile approach for analyzing high-dimensional point clouds. It was applied to cytometric data in this analysis, but it is broadly applicable to a variety of biological and non-biological data sets that can be modeled as high-dimensional point clouds. We have implemented SPADE in MATLAB and the source code is available (**Supplementary Data Set 3**).

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding from National Cancer Institute Integrative Cancer Biology Program (ICBP), grants U56CA112973 and U54CA149145 to S.K.P. A Damon Runyon Cancer Research Foundation Fellowship supports S.C.B. National Science Foundation Graduate Research Fellowship and Stanford DARE Fellowship support K.D.G. This work is also supported by US National Institutes of Health grants U19 AI057229, P01 CA034233, HHSN272200700038C, 1R01CA130826, 5U54 CA143907, RB2-01592, PN2EY018228, N01-HV-00242, HEALTH.2010.1.2-1 (European Commission), as well as the California Institute for Regenerative Medicine (DR1-01477) to G.P.N.

AUTHOR CONTRIBUTIONS

P.Q., G.P.N. and S.K.P. conceived the study and developed the method. E.F.S., S.C.B. and K.D.G.Jr. performed mass and flow cytometry experiments, and participated in the biological interpretation. P.Q., R.V.B., M.D.L. and K.S. performed robustness analysis of the method. P.Q., E.F.S., S.C.B., K.D.G.Jr., G.P.N. and S.K.P. wrote the manuscript and developed the figures.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at. http://www.nature.com/naturebiotechnology.

Published online at http://www.nature.com/nbt/index.html.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Chattopadhyay, P. *et al.* Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry. *Nat. Med.* **12**, 972–977 (2006).
2. Bandura, D.R. *et al.* Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81**, 6813–6822 (2009).
3. Herzenberg, L., Tung, J., Moore, W., Herzenberg, L. & Parks, D. Interpreting flow cytometry data: a guide for the perplexed. *Nat. Immunol.* **7**, 681–685 (2006).
4. Ellis, B., Haaland, P., Hahne, F., Le Meur, N. & Gopalakrishnan, N. Flowcore: basic structures for flow cytometry data. R package version 1.10.0. (2009).
5. Murphy, R.F. Automated identification of subpopulations in flow cytometric list mode data using cluster analysis. *Cytometry* **6**, 302–309 (1985).
6. Lo, K., Brinkman, R. & Gottardo, R. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry A* **73**, 321–332 (2008).
7. Boedigheimer, M. & Ferbas, J. Mixture modeling approach to flow cytometry data. *Cytometry A* **73**, 421–429 (2008).
8. Chan, C. *et al.* Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry A* **73**, 693–701 (2008).
9. Walther, G. *et al.* Automatic clustering of flow cytometry data with density-based merging. *Adv. Bioinformatics*, published online, doi:10.1155/2009/686759 (19 November 2009).
10. Pyne, S. *et al.* Automated high-dimensional flow cytometric data anlysis. *Proc. Natl. Acad. Sci. USA* **106**, 8519–8524 (2009).
11. van Lochem, E.G. *et al.* Immunophenotypic differentiation patterns of normal hematopoiesis in human bone marrow: Reference patterns for age-related changes and disease-induced shifts. *Cytometry B Clin. Cytom.* **60**, 1–13 (2004).
12. Zare, H., Shooshtari, P., Gupta, A. & Brinkman, R. Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics* **11**, 403 (2010).
13. Bagwell, B.C. Probability state models. US patent 7,653,509 (2010).
14. Bendall, S.C. *et al.* Single cell mass cytometry of differential immune and drug responses across the human hematopoietic continuum. *Science* **332**, 687–696 (2011).
15. Fruchterman, T. & Reingold, E. Graph drawing by force-directed placement. *Softw. Pract. Exp.* **21**, 1129–1164 (1991).
16. Bryder, D., Rossi, D. & Weissman, I.L. Hematopoietic stem cells: the paradigmatic tissue specific stem cell. *Am. J. Pathol.* **169**, 338–346 (2006).
17. Chao, M.P., Seita, J. & Weissman, I.L. Establishment of a normal hematopoietic and leukemia stem cell hierarchy. *Cold Spring Harb. Symp. Quant. Biol.* **73**, 439–449 (2008).
18. Ashwell, J.D. The many paths to p38 mitogen-activated protein kinase activation in the immune system. *Nat. Rev. Immunol.* **6**, 532–540 (2006).
19. Guha, M. & Mackman, N. Lps induction of gene expression in human monocytes. *Cell. Signal.* **13**, 85–94 (2001).
20. Chen, W. *et al.* Thrombopoietin cooperates with flt3-ligand in the generation of plasmacytoid dendritic cell precursors from human hematopoietic progenitors. *Blood* **103**, 2547–2553 (2004).
21. Qiu, P., Gentles, A.J. & Plevritis, S.K. Discovering biological progression underlying microarray samples. *PLoS Comput. Biol.* **7**, e1001123 (2011).

ONLINE METHODS

Flow cytometry analysis of mouse bone marrow. Bone marrow was collected from the femurs and tibia of 6- to 10-week-old C57BL/6 mice. Cells were stained for 30 min at 4 °C in FACS buffer (PBS + 0.5% BSA + 0.02% NaN_3). The following markers were used in staining: c-kit, Sca-1, CD150, CD11b, B220, TCRβ, CD4 and CD8. All animal studies were done in compliance with the Stanford Administrative Panel on Laboratory Animal Care Protocol 15986. Data were collected using the Becton-Dickinson LSR2 flow cytometer, and transformed using inverse hyperbolic sine transformation[22]. One initial gate was applied based on forward and side scatters to exclude doublets and debris.

Mass cytometry analysis of human bone marrow. Next-generation mass cytometry data were obtained from reference 14. Briefly, fresh adult healthy whole human bone marrow (BM) was purchased from All Cells, where it was collected under an Institutional Review Board–approved protocol. Ficoll-separated bone marrow mononuclear cells were stimulated using 24 unique perturbation conditions, fixed with paraformaldehyde, stained for surface markers, washed, permeabilized with methanol, stained for intracellular markers, washed, stained with an iridium-tagged DNA intercalator, and then measured on the CyTOF mass cytometer (DVS Sciences).

Overview of SPADE. SPADE is performed in four steps. (i) Density-dependent down-sampling to equalize the density in the point cloud of cells, (ii) agglomerative clustering to partition the point cloud of cells into cell clusters, (iii) minimum spanning tree construction to link the cell clusters and (iv) up-sampling to map all the cells onto the resulting tree structure.

(i) Density-dependent down-sampling. SPADE views a cytometry data set as a high-dimensional point cloud, where each point in the cloud is one cell and the dimension of the cloud is the number of cellular markers. Dense regions of the cloud correspond to abundant cell types, whereas low-density regions correspond to rare cell types or cells in transition between abundant cell types. Most clustering algorithms rely on the density variation to identify abundant cell types[6–10,12]. In contrast, SPADE down-samples the data in a density-dependent fashion to remove the density variation.

SPADE estimates the local density (LD_i) for cell i, defined as the number of cells within its neighborhood. We use an L1 distance metric to compute the distance between cells. The size of the neighborhood is chosen such that most cells have at least one neighbor (see pseudo-code in **Supplementary Section 8**). According to the target density (TD) and outlier density (OD), SPADE keeps each cell i with the following probability:

$$prob(keep\ cell\ i) = \begin{cases} 0, & if\ LD_i \le OD \\ 1, & if\ OD < LD_i \le TD \\ \dfrac{TD}{LD_i}, & if\ LD_i > TD \end{cases}$$

Thus, cells whose local densities are $<OD$ are discarded. Cells whose local densities are between OD and TD are not down-sampled. Cells in high-density regions are heavily down-sampled such that their local densities are reduced to TD. The target density can be defined by the local density of the rare cell types of interest. In the simulated data (**Fig. 1**), we chose OD and TD to be the 1st and 3rd percentiles of the local densities of all the cells. SPADE down-sampled the data from 20,000 cells to ~4,000 cells. Although the size of the data was significantly reduced, most cells of the rare cell type remained after down-sampling, and the shape of the point cloud was preserved.

The purpose of density-dependent down-sampling is to increase the prevalence of rare cells, so that SPADE is able to identify them in the subsequent clustering and tree construction steps. However, down-sampling also increases the prevalence of nonspecific noise events whose local densities are $>OD$. This is a trade-off between signal and noise.

(ii) Agglomerative clustering. SPADE employs a variant of an agglomerative hierarchical clustering algorithm. At the beginning of the first iteration of the agglomerative process, each cell forms its own cluster. One cell is randomly chosen and grouped with its nearest neighbor, defined by single linkage L1 distance. Then, another cell is randomly chosen from the remaining cells and grouped with its nearest neighbor, if the nearest neighbor has not already been grouped with other cells in the current iteration. After all the cells are examined (that is, either chosen or grouped with other cells), the first iteration ends and the number of clusters is reduced by approximately half. The same procedure is repeated in the second iteration to further reduce the number of clusters by approximately half. The iterative process continues until the number of remaining clusters reaches a user-defined threshold. Clustering simplifies the point cloud, distilling it into abutting cell clusters that span the full space occupied by the original cloud. The scale of the simplification can be controlled by adjusting the desired number of clusters.

(iii) Minimum spanning tree construction. SPADE uses Boruvka's algorithm[23] to construct a minimum spanning tree (MST) that links the cell clusters. Each cell cluster is one tree node, and is represented by its median marker expressions. Briefly, we start from a graph with no edges, and iteratively add edges. In each iteration, we randomly select one connected subgraph, calculate its single linkage L1 distances to all nodes outside the randomly selected subgraph, and add an edge that corresponds to the smallest single linkage distance. This process iterates until all nodes are connected. As the MST tends to connect clusters that are close to each other to achieve the minimum total edge length, the resulting tree approximates the shape of the point cloud.

(iv) Up-sampling. To calculate the median intensity and other statistics of each cluster with high accuracy, SPADE performs up-sampling by assigning each cell in the original data set to one cluster. For each cell in the original data set, SPADE finds its nearest neighbor in the down-sampled data (subset of data used in clustering), and assigns this cell to the cluster that the nearest neighbor belongs to.

Visualization of the SPADE tree. SPADE produces the topology of a tree structure. When visualizing the SPADE tree, we can arbitrarily rotate the layout, alter the angles between branches or change the length of the edges. These operations change the appearance of the SPADE tree. However, as long as the topology is not changed, it still represents the same result. To automatically determine a layout of the SPADE tree, we used a modification of the Fruchterman-Reingold algorithm[15]. The layout algorithm works as follows: we first find the longest path in the tree, and fix nodes in the longest path on an arch-like curve. The rest of the tree nodes are gradually appended to the main arch. When a new node and a new edge are appended to the set of nodes that are already visualized, the position of the new node is determined by simulating (i) a repelling force between each existing node and the new node, and (ii) an attracting force generated by the new edge. The simulated physics system is the reason why smaller branches are oriented pointing outwards from the main arch.

Annotation and interpretation of the SPADE tree. After visualizing the SPADE tree and overlaying colors on the tree nodes, we derive annotations manually, according to the colored trees. The boundaries are manually drawn to separate regions that show drastically different colors. Gating and prior knowledge are not used to draw the boundaries. Prior knowledge is used to interpret the biological relevance of each tree region. Although the annotation of the SPADE tree involves a certain level of subjective interpretation, we believe that SPADE is less subjective than gating because the interpretation is guided by the SPADE tree, which encodes an objectively derived topology among all cell types underlying the data. In contrast, gating analysis is entirely guided by the user's prior knowledge, and each gating plot only displays a two-dimensional (2D) subset of the data where even the order that cell populations are gated in can drastically affect the endpoint subsets. SPADE 'see' all the dimensionality that even multiple 2D gating plots miss.

Parameter selection for SPADE analysis. The input parameters of SPADE include: markers used to build the SPADE tree, outlier density, target density and desired number of clusters. The main tuning parameters are the markers to use and the desired number of clusters.

Choice of markers used in SPADE relies on the user's prior knowledge of which markers can be used to organize the cellular heterogeneity underlying the data. This input is important because the shape of the cell cloud may be

different when different sets of markers are used (see **Supplementary Section 2**). Owing to the correlation among protein markers, as long as the majority of selected markers are meaningful, SPADE is robust to exclusion of a few meaningful markers or inclusion of irrelevant ones. In the human bone marrow analysis, even when NK-specific markers were not used, SPADE clustered the NK cells together (**Fig. 4**). In this data set, CD90 did not provide an informative signal but was among the 13 surface markers used by SPADE, and SPADE still produced a meaningful tree.

Outlier density is used to exclude cells with the lowest local densities. If it is set to the 1st percentile of local densities of all the cells, the bottom 1% of cells with lowest local densities are regarded as noise and discarded. Note that such a choice does not necessarily mean that rare stem cells (that is, 0.2% of the population) will be discarded. If the stem cells are similar to each other and form a 'clique', their local densities could be much higher than cells that represent noise. In all our current analyses, we choose outlier density to be the 1st percentile of the local densities.

Target density determines how many cells will survive the down-sampling process. The choice depends, in part, on the density of the rare population that the user aims to detect. Another purpose of this parameter is to reduce the number of cells, so that the subsequent clustering step is computationally more tractable. Ideally, we would like to set the target density comparable to the local density of the rare cells. However, when there is no prior knowledge of which cells are the rare cells, it is difficult to optimize the value of the target density. In the mouse bone marrow analysis, the choice of 5th percentile was empirical. In the human bone marrow analysis, because we were pooling multiple data sets and we wanted different data sets to contribute an equal number of cells, we varied the target density such that a fixed number of 20,000 cells would survive the down-sampling step for each data set. For most of our current analyses, we chose the target density to produce 20,000 cells after down-sampling.

The desired number of clusters determines the stopping criterion of the agglomerative clustering process and the number of nodes in the SPADE tree. If the number of clusters is too small, the SPADE tree cannot correctly capture the shape of the cloud. If this number is too large, the SPADE tree is not easily interpretable. The choice of this parameter depends on the complexity of the shape of the cloud. We suggest that this parameter be set much larger than the number of expected subpopulations in the data. In the mouse and human bone marrow analysis, if we double this parameter, roughly every tree node will be split into two, and the general topology of the resulting tree will remain the same. In our current practice, the desired number of clusters is usually set to be 50, 100 or 300.

SPADE for comparing multiple data sets. SPADE can be used to compare multiple experiments, with overlapping staining panels. After separately down-sampling the data from each individual experiment, we can pool the down-sampled data into a meta-down-sampled data set, which is a meta-cloud that represents where a cell may be in the high-dimensional space defined by the markers that are common across the experiments, that is, the 13 core surface markers in the human bone marrow data set. The SPADE tree represents the shape of the meta-cloud. By coloring the tree using the common markers, we can annotate the tree and sketch out the phenotypic landscape of the meta-cloud. For a marker that varies across experiments, its behavior can be visualized by contrasting its intensities across different experiments. Furthermore, cells in one experiment may not populate the entire meta-cloud. We can color the tree using the change of cell frequencies between difference experiments, which allows us to observe whether any phenotypes emerge or disappear in response to perturbations.

22. Kotecha, N., Krutzik, P.O. & Irish, J.M. Web-based analysis and publication of flow cytometry experiments. *Curr. Prot. Cytom.* **53**, 10.17.1–10.17.24 (2010).
23. Pettie, S. & Ramach, V. An optimal minimum spanning tree algorithm. *JACM* **49**, 49–60 (1999).

doi:10.1038/nbt.1991