Review of the Clustering and Feature Extraction Methods of Mass Cytometry

Yunqin Lee

Introduction

Single cell analysis of heterogeneous populations requires the simultaneous quantitative determination of multiple biomarkers for functional and phenotypical identification. However, detailed characterization of single cells has been limited by available probes and instrumentation, where the number of parameters in conventional flow cytometers has been restricted by spectral overlap between fluorescent probes. Current fluorescence-based flow cytometry technologies typically provide measurements of up to 10-12 parameters, which have been expanded to 17 parameters using improved fluorescent probes or quantum dots with narrower emission bandwidths (Chattopadhyay et al. 2006). On the other hand, limitations in cytometry instrumentation have been overcome with the development of the next-generation mass cytometry platform (CyTOF) (Bandura et al. 2009). The mass cytometer is based on inductively coupled plasma time-of-flight mass spectrometry and currently allows single cell measurement of more than 30 parameters (Bandura etal. 2009, Bendall et al. 2011, Gibbs et al. 2012, Newell et al. 2012).

As a corollary, analysis of multidimensional cytometric data has become progressively more complex with more measured parameters. Despite the high throughput nature of the single cell measurements, the current methods for data analysis remain surprisingly low throughput, requiring manual selection of cell subsets in a labor-intensive and subjective manner (Herzenberg et al. 2006, Qiu et al. 2011). The flow cytometry data are stored in flow cytometry standard (FCS) files and extracted using software such as FlowJo and FlowCore (Ellis et al. 2009). These software allow each data file to be viewed as biaxial plots of two parameters, where cells expressing phenotypic markers of interest are selected for by manually drawing a 'gate' demarcating the boundaries of the cell subset (Herzenberg et al. 2006). Further phenotypic characterization is performed by sequentially 'gating' on biaxial plots featuring other parameters in the downstream analysis. The shape, location, and sequence of the gates depend on the investigator's knowledge of the biological system and interpretation of the experiment, which is subjective and likely to vary with different investigators (Herzenberg et al. 2006). Depending on the manually drawn gates, cell subsets may also be inadvertently excluded, and relationships between unpaired parameters may remain undiscovered (Boedigheimer et al. 2008).

To reduce the inherent user variability and allow a more comprehensive overview of the cell phenotypes, automatic gating algorithms have been independently developed by several groups (Boedigheimer et al. 2008, Chan et al. 2008, Lo et al. 2008, Murphy 1985, Pyne et al. 2009) to provide computer-assisted objective data analysis that are not predicated on user-defined gates. Some of the earliest approaches of automatic gating algorithms utilize k-means clustering (Murphy 1985) and Gaussian mixture modeling (Demers et al. 1992) to identify cell subsets. Mixture modeling assumes that each sample consists of a mixture of components modeled as multivariate distributions, allowing cell subsets to be described using continuous changes in expression in all dimensions, instead of discrete changes in expression (e.g. binary high/low expression) in selected dimensions as determined by manual gating with rigid boundaries. Statistical mixture modeling for flow cytometry is an active area of research, with the development of multiple algorithms. Although Guassian mixture modeling has been refined for flow cytometric analysis (Boedigheimer et al. 2008, Chan et al. 2008), errors in clustering may still arise from outliers and skew in the data. Alternative algorithms have been developed to transform the outliers and data skews, thus reducing clustering errors (Lo et al. 2008, Pyne et al. 2009).

However, one common limitation across these algorithms is the inability to detect rare cell types, which are often excluded as outliers or absorbed into larger clusters. Recent algorithms have started to include mechanisms for identification of rare events, such as SamSPECTRAL (Zare et al. 2010), which uses a data reduction scheme to down-sample abundant cell populations using potential theory, allowing detection of populations comprising between 0.2% to 2% of the total data. A second limitation is the inability of these algorithms to detect intermediate phenotypes that are typical of the continuous progression of cellular differentiation in heterogeneous samples (van Lochem et al. 2004). A third limitation is the scalability issue of visualizing increasing numbers of parameters per cell. Biaxial plots display the correlation of two parameters only, such that visualization of *m* parameters would require a total of $m(m-1)/2$ biaxial plots. Identifying correlations and relationships in high-dimensional data from a series of biaxial plots becomes a tedious and labor-intensive process. One approach to this issue is the probability state model, implemented in the Gemstone software package (Bagwell 2010), where cells are rearranged linearly according to a predetermined expression pattern of the parameters. However, this semi-supervised approach requires knowledge of the progression of marker expression underlying the cell populations, and does not allow for branching during the rearrangement (Qiu et al. 2011).

To address the aforementioned limitations, spanning-tree progression analysis of density-normalized events (SPADE) was developed to complement the technique of mass cytometry (Qiu et al. 2011). SPADE is an unsupervised algorithm that analyzes high-dimensional mass cytometry data and objectively organizes cells into a hierarchy of related phenotypes without any prior knowledge. The SPADE algorithm is able to identify rare populations, such as hematopoietic stem cells (Bendall et al. 2011, Qiu et al. 2011), and enable visualization of multiple cell types in a single branched "minimum-spanning tree" structure. The clustering and feature extraction methods of mass cytometry will be reviewed in this article.

CyTOF Technology

The instrumentation for mass cytometry was developed by Bandura et al., and is based on inductively coupled plasma time-of-flight mass spectrometry.

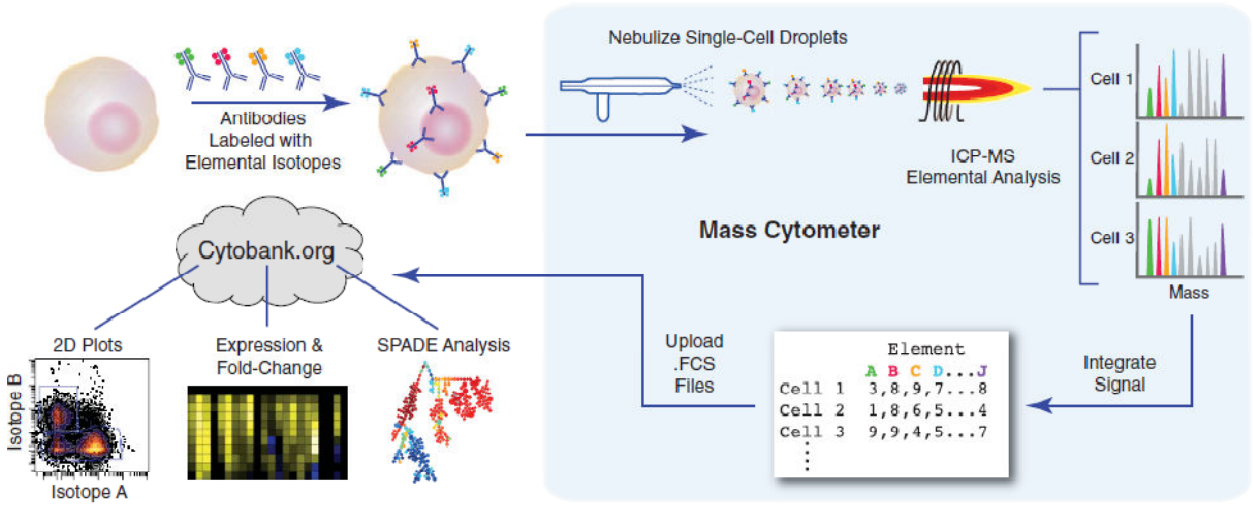

Figure 1: Workflow summary of mass cytometry analysis. (Bendall et al. 2011)

The sample is prepared by staining cells with epitope-specific antibodies conjugated to transition element isotopes of unique mass (instead of fluorescent labels). The rare earth elemental tags (such as lanthanides) are attached to antibodies through metal-chelator coupling reagents. The sample is then

introduced into the mass cytometer, where the individual cells undergo the processes of vaporization, atomization, and ionization by inductively coupled plasma (ICP). The elemental isotopes are detected by a time-of-flight mass spectrometer (TOF-MS). In other words, the cells are nebulized into single-cell droplets of elemental tags which are read to generate an elemental mass spectrum. Hence, mass cytometry uses mass, rather than conventional fluorescence, as the readout. The integrated elemental reporter signals for the cells can then be analyzed using various approaches, such as traditional biaxial plots, heat maps of changes in expression, and tree plots (Bandura et al. 2009, Bendall et al. 2011).

Mass cytometry utilizes the high resolution, sensitivity, and speed of analysis of ICP-TOF-MS to achieve simultaneous measurements of multiple protein markers (Bandura et al. 2009). Mass cytometry overcomes the key limitation of multi-parameter single-cell measurement by eliminating the dimensional restriction caused by spectral overlap of fluorescent labels, hence theoretically allowing simultaneous measurements of up to 50 parameters or more. Another advantage of mass cytometry is the precision of each measurement generated from mass spectrometry, where detection of each elemental tag is a discrete event with no overlaps between the detection channels, thus eliminating the additional step of signal compensation (which is otherwise required for traditional fluorescence-based flow cytometry). Rare earth elemental isotopes are also not naturally found in biological systems, thus removing the need for background correction.

However, the main disadvantage is that the mass cytometer is incapable of cell sorting, as cells are completely nebulized during the measurement (Bandura et al. 2009, Bendall et al. 2011). Mass cytometry also has a lower sampling efficiency of less than 30% (compared to 95% for fluorescence-based flow cytometry), lower sampling rate of 2 million cells per hour (compared to 25-60 million per hour for flow cytometry) (Bendall et al. 2012). Another challenge is the optimization process of procuring and testing a panel of antibodies, especially as monoclonal antibodies have different affinities, stabilities, and resistance to conjugation chemistries (Bandura et al. 2009).

Methods - SPADE

In SPADE, fundamentally, the data is analyzed as a high-dimensional point cloud of cells, and underlying patterns and geometry of the data are inferred using topological methods (Qiu et al. 2011). SPADE comprises four computational modules; density-dependent down-sampling, agglomerative clustering, construction of a minimum spanning tree, and up-sampling (Figure 2).

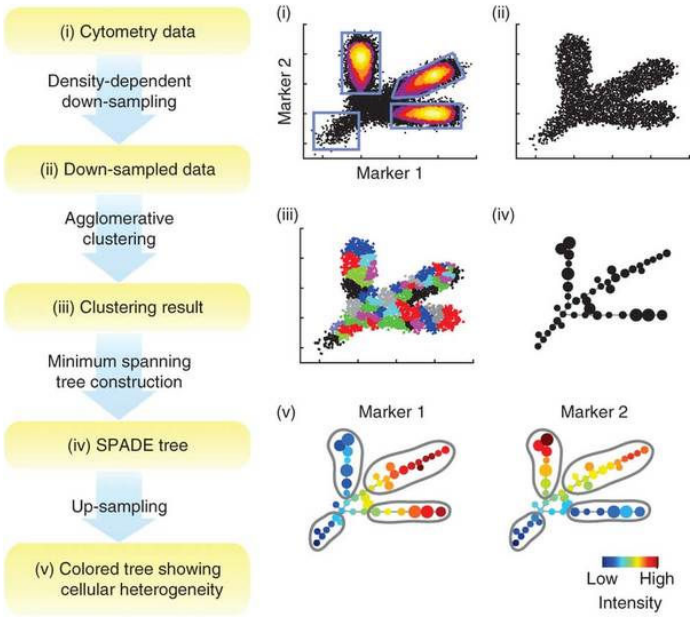

Figure 2: Flowchart of SPADE analysis of a simulated two-parameter data set, with one rare population and three abundant populations (Qiu et al. 2011).

1. Density-dependent down-sampling

The cytometry data set is analyzed as a high-dimensional point cloud, where each cell is represented as a point in the cloud, each marker is represented by a dimension of the cloud (Qiu et al. 2011). Cloud density would then directly reflect the abundance of a specific cell type, such that rare cell types would be found in regions of low densities,

while abundant cell types would be found in regions of high densities. The density variation is removed by density-dependent down-sampling to equalize the density of the cloud regions, and to represent rare and abundant cell types equally in subsequent analyses (Qiu et al. 2011).

Density-dependent down-sampling for each cell was performed according to the computed local density, target density, and outlier density. The local density for each cell was defined as the number of cells within a neighborhood as defined by a distance threshold, which was in turn defined as a multiple of the median minimum distance between a cell and its nearest neighbor. A L1 (Manhattan) distance metric was used where the distance between two cells is the sum of the absolute differences of their parameter values. This allowed most cells to have at least one neighbor within the distance threshold. On the other hand, the outlier density and target density are user-specified inputs for SPADE, usually determined empirically.

Outlier density is used to exclude the most phenotypically isolated cells with the lowest local densities. The current default for SPADE analyses is to set the outlier density as the 1st percentile of local densities of all cells, such that the bottom 1% of the cells with the lowest local densities are regarded as noise and discarded. This does not result in the exclusion of rare cell types, such as hematopoietic stem cells that make up 0.2% of the bone marrow population, as phenotypically similar stem cells may form clusters of high local densities in the point cloud.

Target density determines the degree of down-sampling performed on the original data set. Ideally, the target density selected should be comparable to the local density of the rare population of interest. However, it is difficult to optimize the value of the target density without knowledge of the positions of the rare cell types in the point cloud. The current default for SPADE analyses is to set the target density as the percentile that would produce 20,000 cells after down-sampling (Qiu et al. 2011). The default target of 20,000 down-sampled cells is to make the subsequent clustering step more computationally tractable.

Down-sampling was performed by computing the probability of retaining a cell *i* as determined by the local density (LD_i), target density (TD), and outlier density (OD) (Qiu et al. 2011):

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

Cells with local densities lower than the outlier densities are excluded, while cells with local densities between the outlier and target densities are retained. Cells with local densities higher than the target densities (i.e. in high-density regions) are down-sampled such that their local densities are reduced to the target density. Hence the extent of down-sampling is dependent on the density of the cells.

A major advantage of down-sampling is the equal representation of rare and abundant cell types. As most of the points belonging to rare cell types are retained, the rare cells are able to form their own clusters without being outnumbered by the abundant cell types in the subsequent analysis. The overall shape of the point cloud is preserved in the process. The size of the dataset is also significantly reduced, making subsequent analyses more computationally tractable. However, one disadvantage is the potential overrepresentation of noise events, as nonspecific noise events with local densities higher than the outlier density are retained. The signal-to-noise ratio in the down-sampled dataset may be

reduced compared to the raw dataset, which may affect the clustering methods used to the construct the cellular hierarchy. Thus the selection of appropriate values for the target density and outlier density is critical to ensure maximal representation of rare cell types with minimal inclusion of noise events.

2. Agglomerative clustering

Agglomerative hierarchical clustering is performed to segment the down-sampled point cloud of cells into cluster of cells with similar intensities of the markers (i.e. similar phenotypes). The clustering algorithm is initialized by setting each individual cell as a cell cluster. In each iteration of the algorithm, one cell cluster (equivalent to a single cell for the first iteration) is randomly selected and grouped with its nearest neighbor, as defined by the single linkage L1 distance. Another cell cluster is then randomly chosen from the remaining clusters and grouped with its nearest neighbor, if the nearest neighbor has not been paired with another cluster in the current iteration. Each iteration of the algorithm will result in the pairing of all the cell clusters, reducing the total number of clusters by approximately half. The clustering algorithm performs multiple iterations until the number of cell clusters reaches a user-specified threshold.

This 'bottom up' approach merges cell clusters in a greedy manner that ensures the local optimal solution. One advantage of agglomerative clustering is its speed, which is a nontrivial matter when dealing with datasets containing large numbers of cells. The resolution of the clusters and resulting SPADE tree can also be modified by changing the user-specified final desired number of cell clusters (i.e. nodes in the constructed MST) (Qiu et al. 2011). Over-clustering results in too few nodes in the MST, leading to an inaccurate representation of the point cloud. Under-clustering results in too many nodes in the MST, leading to a complex SPADE tree that is not easily interpretable. The choice of the number of clusters depends on the number of markers used for the cytometry experiment and the complexity of the shape of the point cloud. Current implementations of SPADE analyses set the number of clusters to be 50, 100, or 300 (Qiu et al. 2011).

3. Construction of a minimum spanning tree

A minimum spanning tree (MST) was constructed using Boruvka's algorithm (Pettie et al. 1999) to link the cell clusters, summarizing and extracting the topology of the point cloud. Each cell cluster is represented as a tree node with its median parameter values. The graph is initialized with all the tree nodes and no edges. In each iteration of the algorithm, one connected subgraph (equivalent to a single tree node in the first iteration) is randomly selected and all single linkage L1 distances to all nodes outside the randomly selected subgraph are calculated. An edge corresponding to the smallest linkage distance is added to the graph. This process undergoes multiple iterations until all the nodes are connected in a spanning tree with minimum total edge length (Qiu et al. 2011). The resulting MST will resemble the shape of the point cloud, and can be thought of as the topological skeleton of point cloud. Since each edge has a distinct weight, it is likely that only one unique solution for the MST will exist for each data set.

4. Up-sampling

Up-sampling is performed to map all the cells onto the constructed MST structure, allowing the properties of each cell cluster (such as median intensity) to be calculated with higher accuracy. For each cell in the original dataset, the distances to all the cells in the down-sampled dataset are computed to find its nearest neighbor. The cell is then assigned to the cell cluster to which its nearest neighbor belongs.

Force-directed layout of the SPADE tree

The MST tree recapitulates the topology of the point cloud in the form of an unrooted tree formed by hierarchical clustering. A MST tree with a fixed topology can be represented in multiple ways, by rotating the layout, or changing the length of the edges or the angles between the branches (Figure 5a-b). To standardize the layout of the SPADE tree, a modification of the Fruchterman-Reingold algorithm (Fruchterman et al. 1991) for graph drawing was used to automatically determine a layout of the SPADE tree.

For the layout algorithm, the longest path in the MST tree is identified and represented as an arch-like curve. The nodes found on the longest path are fixed onto this main arch, and the remaining tree nodes are appended to this main arch. The position of a new node is determined by two factors: a repelling force between the new node and each fixed node in the layout, and an attracting force between the new edge between the new node and the fixed nodes (Qiu et al. 2011). This algorithm ensures that the nodes are arranged in a manner that would produce in a layout with minimum energy. The force-directed placement of the nodes produces the characteristic structure of the SPADE tree, where smaller branches radiate outwards from a main arch (Figure 3b, Figure 4).

Manual annotation and interpretation of the SPADE tree

The tree nodes are colored according to the median intensities of their parameters, generating colored SPADE trees representative of the segmented point cloud (Figures 3c-e). The user then has to analyze the colored nodes and annotate the SPADE tree manually.

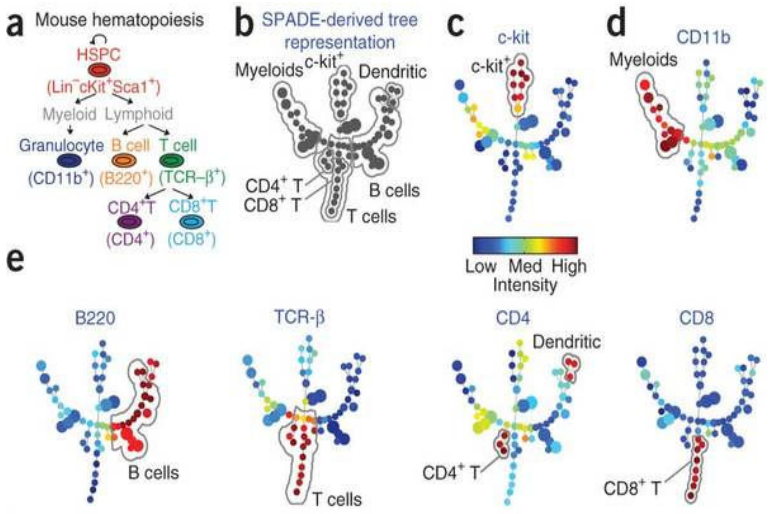

Figure 3: a) Known hematopoietic hierachy in mouse bone marrow. b) SPADE tree derived from mouse bone marrow data manually annotated by the user. (c-e) SPADE trees colored by the median intensity of one individual marker. (Qiu et al. 2011)

Boundaries are manually drawn to separate regions that show different colors (Figures 3c-e). Although gating and prior knowledge are not necessary for drawing the boundaries, prior knowledge is necessary to interpret the biological relevance of each tree region. For example, the user has to be aware that myeloid cells are CD11b+ (Figure 3d), B-cells are B220+, T-cells are TCR-β+ and can be CD4+ and/or CD8+ (Figure 3e) (Bryder et al. 2006).

Cell clusters in the bounded regions are annotated according to their biological relevance, producing an annotated tree representation (Figure 3b, Figure 5). Although manual user-defined annotations inherently involve a degree of subjectivity, they are still less subjective than conventional gating methods. This is because the interpretations and annotations are guided by the topology of the SPADE tree, which is constructed objectively to reflect the phenotypes of the underlying cell subsets in the data (Qiu et al. 2011). Also, SPADE annotations are performed on the entire dataset simultaneously, while conventional gating methods performs annotations on biaxial plots sequentially. Hence the annotated SPADE tree allows visualization of multiple parameters of the entire dataset in a single diagram.

Potential use of a cell ontology for automatic annotation and interpretation of the SPADE tree

The annotation process can be improved with the construction of an ontology of immune cells. Immune cell subsets are defined by the combinatorial expression (or absence) or specific cell markers (Bryder et al. 2006, Chao et al. 2008). These include clusters of differentiation (CD) which are used to label cell surface molecules that are used for immunophenotyping of cells. A cell ontology containing relationships between each cell type and protein biomarkers would provide a context or reference hierarchy by which the annotations may be performed. The cell ontology could also be used to direct the construction of the MST in the SPADE algorithm, by modulating the weights of the edges connecting the tree nodes with information from the cell ontology.

However, although there are efforts in developing cell type ontologies (http://bioportal.bioontology.org/ontologies/1006), current ontologies do not contain the relevant information (e.g. phenotypic expression of cellular markers) for annotating cytometry data. Also, there is no gold standard for selection of markers used to define cell subsets. In fact, selection of markers for immunophenotyping often depends on the conventional markers used in previous published studies or the investigator's preferences (Maecker et al. 2012).

The Human Immunology Project is one example of ongoing efforts towards standardizing immunophenotyping in the human immune system (Maecker et al. 2012). The Human Immunology Project proposes the use of precise and standardized assays to distinguish true biological changes form technical artefacts.

Fluorochrome	Marker				
	T cells	T_{Reg} cells	T_H1, T_H2 and T_H17 cells	B cells	DCs, monocytes and NK cells
FITC	Live or dead	Live or dead	Live or dead	Live or dead	Live or dead
PE	CCR7	CD25	CXCR3	CD24	CD56
PerCP-Cy5.5	CD4	CD4	CD4	CD19	CD123
PE-Cy7	CD45RA	CCR4	CCR6	CD27	CD11c
APC	CD38	CD127	CD38	CD38	CD16
APC-H7	CD8	CD45RO	CD8	CD20	CD3, CD19 and CD20
V450	CD3	CD3	CD3	CD3	CD14
V500	HLA-DR	HLA-DR	HLA-DR	IgD	HLA-DR

Figure 4: Eight-color antibody panels proposed by the Human Immunophenotyping Constortium (Maecker et al. 2012).

Eight-color antibody panels for fluorescence-based flow cytometry have been proposed for the immunophenotyping of specific immune cell subsets (Figure 4). It is conceivable that the protein markers selected for the panels could serve as the beginnings of an immune cell ontology, where description logic can be used to describe the specific attributes for a cell subset. Possible attributes include the expression or lack of protein markers, their parent cell subsets, and their differentiated progeny cell subsets. The immune cell ontology can be created using ontology editors such as Protégé-OWL (http://protege.stanford.edu). The use of description logic and OWL also allows the automatic generation of an inferred hierarchy of the cell subsets using reasoners built into Protégé. The establishment of this immune cell ontology would be indispensable for automating the annotating of the SPADE tree, and refine the algorithm used in constructing the MST tree.

Marker selection for SPADE analysis

The user-determined input parameters for SPADE include the markers selected to build the SPADE tree, the outlier density and target density for the down-sampling algorithm, and the desired number of clusters for the agglomerative clustering algorithm.

Selection of the appropriate markers to be used in the mass cytometry experiment is critical for the accurate elucidation of the underlying cellular hierarchy. This is because the shape of the cell cloud changes when different sets of markers are used, resulting in different SPADE trees being generated. Due to correlation amongst protein markers, it has been shown that SPADE analysis is robust to the exclusion of a few meaningful markers or inclusion of a few irrelevant markers, given that the majority of the selected markers are meaningful (i.e. they are largely sufficient for differentiating between cell types) (Qiu et al. 2011). Selection of markers has been performed using the investigator's prior knowledge of the biological system (Bendall et al. 2011, Gibbs et al. 2012, Newell et al. 2012). However, this could be performed using standardized markers from the Human Immunophenotyping Project (Maecker et al. 2012), or potentially, by using an immune cell ontology.

Comparing multiple datasets with SPADE

SPADE can be used to compare multiple cytometry datasets with overlapping staining panels of antibodies. Down-sampled data from each individual dataset is pooled to form a meta-down-sampled dataset depicting all the cells in the point cloud space according to the markers in common across all the datasets; e.g. 13 core surface markers for human bone marrow (Bendall et al. 2011, Qiu et al. 2011). The resulting SPADE tree reflects the phenotypic topology of the meta-dataset, and can be annotated manually (Figure 5a-b).

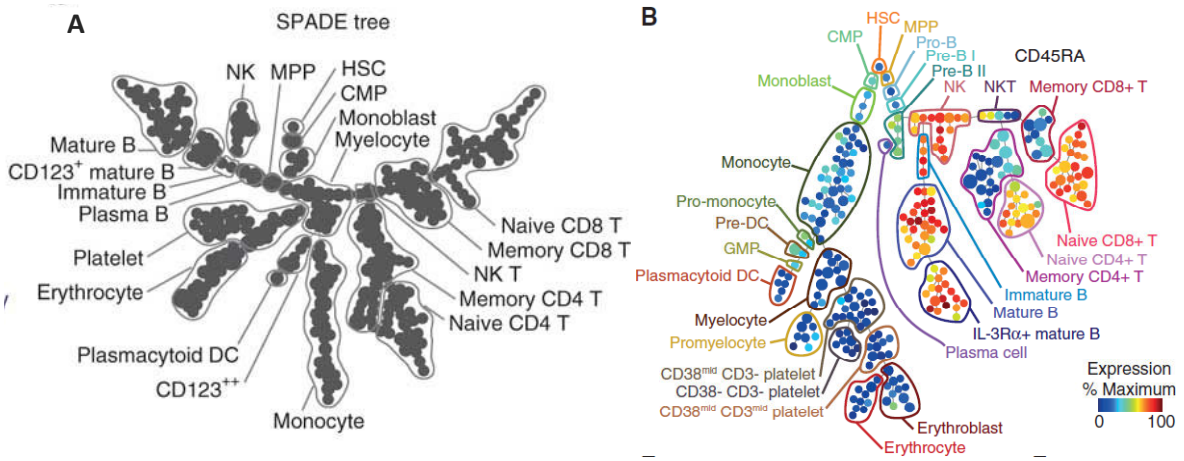

Figure 5: Annotated SPADE trees showing immunophenotypic progression in human bone marrow using 13 cell-surface markers. A) Layout automatically generated by the Fruchterman-Reingold algorithm (Qiu et al. 2011). B) Layout manually reorganized to resemble the classic immunology diagram of hematopoietic developmental hierachy (Bendall et al. 2011).

The behavior of a marker in response to a perturbation can be visualized by displaying the ratio or difference of intensities between the stimulated and unstimulated (basal) conditions for each node, overlaid onto the annotated SPADE tree. This allows direct observation of all cell subsets in the hematopoietic compartment that were affected by the perturbation (Bendall et al. 2011).

Methods – Principal Component Analysis

While SPADE analysis is effective in showing relationships between different cell types and discovering potential new cell types, one limitation of using SPADE clustering analysis is that the algorithm emphasizes the similarities between cells and the assignment of distinct cell types by the algorithm is arbitrarily determined by the down-sampling and up-sampling processes.

Another statistical method used to visualize multidimensional single cell data is principal component analysis (PCA) (Pearson 1901). PCA uses an orthogonal transformation to collapse the dataset containing correlated parameters to a smaller set of linearly uncorrelated variables known as principal components (PC), such that each principal component is a weighted combination of all the markers. The principal components are arranged in descending order of the variance, i.e. PC1 has the largest possible variance (Haining 2012).

In a study by Newell et al., mass cytometry was used to determine patterns of cytokine expression and virus-specific cell niches within a continuum of CD8+ T-cell phenotypes (Newell et al. 2012). PCA was used to visualize the information from 25 functional and phenotypic markers on PMA-ionomycin stimulated CD8+ T-cells, where each principal component was described as a weighted combination of the 25 markers.

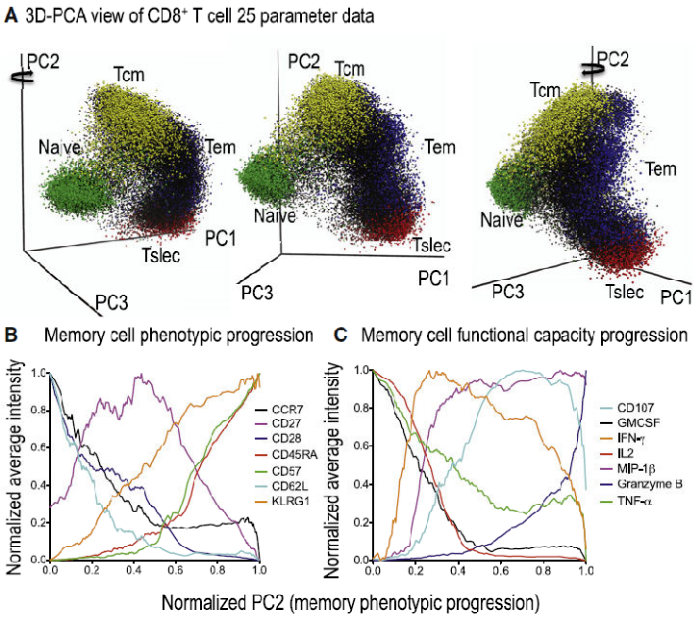

Figure 6: 3D-PCA representation of CD8 T-cell 25 parameter data and memory cell phenotypic and functional progression (Newell et al. 2012).

The first three principal components (PC1, PC2, PC3) of the CD8+ T-cell compartment were selected and visualized in three dimensions, using the protein structure program PyMol (DeLano, 2002). Naïve, central memory (Tcm), effector memory (Tem), and short-lived effector (Tsle) CD8+ T-cells were manually gated in biaxial plots based on stringent surface expression criteria and differentially colored on the PCA plot (Newell et al. 2012) (Figure 6a). A folded Y-shape pattern was observed with naïve T-cells at the base of the Y and Tsle and Tcm cells forming distinct nodes at the tips of the Y. The arrangement of the subsets suggests a continuum of phenotypes connecting the naïve and memory subsets, and a continuum of memory cells connecting Tcm cells to Tem cells, which were then connected to Tsle cells (Newell et al. 2012). The graded variations in the functional and phenotypic markers associated with memory cell progression along PC2 also reflected known patterns of protein expression during memory cell differentiation (Figure 6b-c).

The advantage of using 3D-PCA is that no distinct cell clusters were arbitrarily created (unlike SPADE); instead, all the cells were individually represented on the 3D plot such that the underlying graded progression between the clusters were easily visualized. This approach allows the identification of major cell subsets defined by selected markers without any biases. The overall phenotypic and functional characteristics of each perceived cell cluster and their interrelatedness can also be perceived intuitively by viewing the 3D-PCA plot. The main disadvantage of PCA is the loss of the ability to determine the expression of individual markers by each cell.

Conclusion

The development of mass cytometry has allowed the simultaneous measurement of more than 50 parameters on single cells (Bandura et al. 2009). This marked increase in the dimensionality of single-cell data has resulted in a need for novel methods of analyzing cytometry data. Although automatic gating

methods for identifying cell subsets have been previously proposed by multiple groups (Boedigheimer et al. 2008, Chan et al. 2008, Lo et al. 2008, Murphy 1985, Pyne et al. 2009), visualization of flow cytometry data has remained, in essence, as multiple series of biaxial plots.

One elegant method used for analyzing and visualizing these large datasets is SPADE (Qiu et al. 2011), which reduces a high-dimensional data set to an intuitive tree diagram that reflects the relationships between cell types. SPADE has been shown to be able to recapitulate known patterns of hematopoietic differentiation with much finer granularity (Bendall et al. 2011), demonstrating its validity as a clustering algorithm for multi-dimensional data. Mass cytometry and SPADE have also been used in other contexts besides phenotyping cells of the immune compartment, such as the potential identification of tumor-initiating cells in acute myeloid leukemia (Gibbs et al. 2012).

However, the use of SPADE analysis requires finding the optimal set of input parameters such as the markers for analysis, target and outlier density, and the desired number of cell clusters in the SPADE tree. The construction of the SPADE tree could be facilitated with the development of an immune cell ontology, which will refine the connecting edges between nodes and allow guided annotations of the resulting cell clusters in SPADE. Alternative methods include using PCA to identify major cell clusters without any manipulation of the position of each individual cell. However, PCA removes the ability of the user to determine expression of individual markers on each cell.

The increase in the number of measurement parameters result in increasing numbers of different possible combinations of cell phenotypes. Given the graded progression of cell subsets shown by PCA (Newell et al. 2012), it is likely that more heterogeneity will be discovered in the immune cell compartment. However the biological significance of these cellular heterogeneities identified by mass cytometry remains to be discovered. For example, finer cell clusters in a single cell type are often obtained in SPADE, and it remains unknown if these finer cell clusters are an artifact of the clustering algorithm, or if they are functionally different subsets that are reproducible. The challenge for experimental investigators will be to determine if the nodes identified by the clustering and feature extraction methods of cytometry are biologically unique.

References

Bandura et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal Chem.* 2009 Aug 15;81(16):6813-22.

Bagwell, B.C. Probability state models. US patent 7,653,509 (2010).

Bendall et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*. 2011 May 6;332(6030):687-96.

Bendall et al. A deep profiler's guide to cytometry. *Trends Immunol*. 2012 Apr 2.

Boedigheimer et al. Mixture modeling approach to flow cytometry data. *Cytometry A.* 2008 May;73(5):421-9.

Bryder et al. Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. *Am J Pathol.* 2006 Aug;169(2):338-46.

Chan et al. Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry A.* 2008 Aug;73(8):693-701.

Chattopadhyay et al. Quantum dot semiconductor nanocrystals for immunophenotyping by polychromatic flow cytometry. *Nat Med.* 2006 Aug;12(8):972-7.

DeLano WL. Unraveling hot spots in binding interfaces: progress and challenges. *Curr Opin Struct Biol.* 2002 Feb;12(1):14-20.

Demers et al. Analyzing multivariate flow cytometric data in aquatic sciences. *Cytometry*. 1992;13(3):291-8.

Ellis et al. Flowcore: basic structures for flow cytometry data. R package version 1.10.0. (2009).

Fruchterman, T. & Reingold, E. Graph drawing by force-directed placement. *Softw. Pract. Exp.* 1991; 21, 1129–1164.

Gibbs et al. Decoupling of tumor-initiating activity from stable immunophenotype in HoxA9-Meis1-driven AML. *Cell Stem Cell*. 2012 Feb 3;10(2):210-7.

Haining WN. The numerology of T cell functional diversity. *Immunity*. 2012 Jan 27;36(1):10-2.

Herzenberg et al. Interpreting flow cytometry data: a guide for the perplexed. *Nat Immunol.* 2006 Jul;7(7):681-5.

Lo et al. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry A.* 2008 Apr;73(4):321-32.

Maecker et al. Standardizing immunophenotyping for the Human Immunology Project. *Nat Rev Immunol.* 2012 Feb 17;12(3):191-200.

Murphy RF. Automated identification of subpopulations in flow cytometric list mode data using cluster analysis. *Cytometry.* 1985 Jul;6(4):302-9.

Newell et al. Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8+ T cell phenotypes. *Immunity*. 2012 Jan 27;36(1):142-52.

Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine.* 1901; 2 (6): 559–572.

Pettie, S. & Ramach, V. An optimal minimum spanning tree algorithm. *JACM*. 1999; 49, 49–60.

Pyne et al. Automated high-dimensional flow cytometric data analysis. *Proc Natl Acad Sci U S A.* 2009 May 26;106(21):8519-24.

Qiu et al. Discovering biological progression underlying microarray samples. *PLoS Comput Biol.* 2011 Apr;7(4):e1001123.

Qiu et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol.* 2011 Oct 2;29(10):886-91.

van Lochem 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.* 2004 Jul;60(1):1-13.

Zare et al. Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics*. 2010 Jul 28;11:403.