Marker Enrichment Modeling (MEM) for automated cell population characterization and identification in complex tissue micro-environments

VANDERBILT UNIVERSITY
CTTC Center for Technology Transfer & Commercialization

Summary

Marker enrichment modeling (MEM) provides a crucial missing piece for true machine learning analysis of cell identities and phenotypes in complex tissue microenvironments, including human immune disorders and cancer.

Addressed Need

A critical emerging need in biology and clinical research is to automatically group cells by phenotype and characterize their identity based on measured features. Historically, this has been done with a labor intensive manual process called "gating" followed by human expert analysis. Tools from the field of machine learning have recently solved the first part of the problem by automatically grouping cells into clusters based on multidimensional phenotype (i.e. automatic gating). However, the remaining problem of characterizing and identifying automatically discovered cell populations has remained an unsolved challenge for computational tools. Routine tasks for immunologists, such as identifying CD4+ T cells, remain major challenges for computers. Currently, human experts extensively review cell subsets after gating in a time consuming process that is inconsistent from person to person. We recently reviewed this field, generalized cytometry data analysis as 14 Steps, and highlighted the need for true machine learning of cell type at Step 13 ("Learn cell identity"), which currently has no alternative other than human effort (Figure 1 & Diggins et al., *Methods* 2015).

Figure 1: Marker enrichment modeling (MEM) addresses a critical gap in the field of automated cytometry: the need for machines to learn the identity of cell subsets. *MEM creates quantitative descriptions (labels) that are unbiased descriptions of the key features that make each cell subset unique. MEM addresses a key gap in the field at "Step 13" in data analysis ("Learn cell identity") where computers have historically lacked a quantitative language and framework to compare the enriched features of cell subsets in order to make quantitative assessments of cell identity. MEM can also be used to describe tissues and patients based on heterogeneous cell subsets. As input, MEM can use samples or cell subsets identified either by human experts or by computational tools. We expect MEM to be heavily used in teaching computers to identify known and new cell types in clinical research and diagnostic applications (see list). Adapted from Diggins et al., Methods 2015.*

Diggins et al., METHODS-D-13-00272

Table 1 – A modular machine learning workflow for semi-supervised high-dimensional single cell data analysis

Analysis step		Traditional	Additional methods[§]	Method here
Data collection	1) Panel design	Human expert	-	-
	2) Data collection	Human expert	-	-
Data processing	3) Cell event parsing	Instrument software	Bead normalization and event parsing [39]	-
	4) Scale transformation	Human expert	Logicle [47]	-
Distinguishing initial populations	5) Live single cell gating 6) Focal population gating	Biaxial gating + human expert	No event restriction, AutoGate [63]	viSNE + human expert (Figure 1)[†]
Revealing cell subsets	7) Select features	Human expert	Statistical threshold [53]	Human expert[†]
	8) Reduce dimensions or transform data	N/A	Heat plots [64], SPADE [12], t-SNE [65], viSNE [9], ISOMAP [27], LLE [29], PCA in R/flowCore [66]	SPADE[†], viSNE
	9) Identify clusters of cells	Human expert	SPADE, k-medians, R/flowCore, flowSOM [67], Misty Mountain [13], JCM [30], ACCSENSE [68], DensVM [28], AutoGate, *Citrus* [14]	SPADE (Figure 2)[†], viSNE + human expert (Figure 1)
	10) Cluster refinement	Human expert	Citrus, DensVM, R/flowCore	-
Characterizing cell subsets	11) Feature comparison	Select biaxial single cell views	viSNE, SPADE, Heatmaps [25, 53], Histogram overlays [25, 53], Violin or box and whiskers plots [66], Wanderlust [31], Gemstone	Heatmaps (Figure 3A)[†], viSNE (Figure 1C), SPADE (Figure 2C)
	12) Model populations	N/A	Median [53], JCM, PCA	-
	13) Learn cell identity	Human expert	-	Human expert[†] (Figure 1B, Figure 2B, and Figure 3B)
	14) Statistical testing	Prism, Excel	R/flowCore	-

[§]Methods with broad application (e.g. R/flowCore) are listed minimally at select steps based on particular strengths or published applications.
[†]Denotes the primary approach used at each step in the sequential analysis workflow shown here.

Marker Enrichment Modeling (MEM) for automated cell population characterization and identification in complex tissue micro-environments

VANDERBILT UNIVERSITY
CTTC Center for Technology Transfer & Commercialization

Technology Description

We have addressed this critical need for automated cell population identification by creating a set of tools and algorithms we collectively call “marker enrichment modeling” (MEM). MEM is complementary to existing tools and approaches, including expert analysis by humans, SPADE, viSNE, SCAFFOLD, Phenograph, Citrus, and R/flowCore. MEM can work with automatically identified cell subsets (generally a key output of existing approaches) as its input and provides a new type of description of cell subsets that can be read by humans and machines. This is the key missing piece to train computers to achieve important, unsolved machine learning tasks, such as identifying CD4+ T cells or going beyond this and determining whether a population of cells represents cancer cells or healthy cells.

Technology Applications

MEM can be used for the following example applications (and more):

- Characterizing known and unknown cell types automatically

  Examples: characterizing newly discovered cancer cell populations and determining which healthy cells they most resemble; identifying the cell subsets discovered in healthy tissue by cytomic approaches.

- Tracking changes in cell subsets in complex human tissues

  Examples: Identifying features enriched on cells during development; characterizing one population of immune cells residing in different organs or tissues.

- Cytometry quality control

  Example: Determining whether cells from one day's preparation are equivalent to cells from another day's preparation.

- Signaling network analysis

  Examples: Identifying which elements of the signaling network are most strengthened or weakened in a cell subset.

- Precision medicine & patient stratification

  Examples: Classifying patients according to the enriched features of identified cell subsets.

- Monitoring clinical correlates, identifying cellular biomarkers

  Examples: Identifying biomarkers of cells associated with patient outcomes; describing how cell subsets in human tissue change over time with or without a clinical intervention.

- Optimizing marker sets for cell isolation

  Examples: Determining which surface markers are most specific to cells of interest in order to sort them by FACS.

- Ranking key enriched features of cell subsets

  Examples: Identifying subset-enriched features in any scientific data set, such as gene analyses or image recognition

Technology Development Status

Software implementation of MEM has been completed.

CTTC Contact:
Masood A Machingal, Ph.D.
615.343. 3548
masood.machingal@vanderbilt.edu

Vanderbilt Lead Inventors:
Jonathan M Irish, Ph.D.
Kirsten E Diggins
https://my.vanderbilt.edu/irishlab/

VU Reference Number: VU14153