Methods for Discovery & Characterization of Cell Subsets

Jonathan Irish, Ph.D.
Vanderbilt University, Nashville, TN
Assistant Professor of Cancer Biology & Pathology, Microbiology & Immunology

Australasian Cytometry Society
10 October 2015
## Disclosures for Jonathan Irish, Vanderbilt University

<table>
<thead>
<tr>
<th>Role</th>
<th>Companies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-founder &amp; board</td>
<td>Cytobank</td>
</tr>
<tr>
<td>Clinical research</td>
<td>Incyte, Karyopharm</td>
</tr>
<tr>
<td>Speaking honorarium</td>
<td>Novartis</td>
</tr>
<tr>
<td>Invited speaker</td>
<td>Fluidigm</td>
</tr>
</tbody>
</table>
The Big Idea: Automatically Identify All Cell Types in Primary Tissues, Create Reference Models to Study Impact of Disease, Genetic Changes, etc.

Based on Becher et al., *Nature Immunology* 2014
Tools from Machine Learning + High Content Data: Comprehensive, Automatic Mapping of Cell Types

Classical map of the ‘myeloid cell system’

Modern map, computationally generated

Irish, *Nature Immunology* 2014
Based on Becher et al., *Nature Immunology* 2014
Effective data analysis is critical to successful cytometry.
We Now Make Billions of Multi-D Single Cell Measurements

=> Need for Machine Learning Tools & Human Readable Views

35+ dimensional single cell data:

Even if we look through all the 2D plots, multidimensional relationships are still hidden...
Unsupervised Analysis: Not Using Prior Knowledge To Guide the Analysis

Prior knowledge examples: Stem cells express CD34, AML cells express CD45

Supervised Approaches
• Expert gating
• Gemstone
• Wanderlust
• Citrus

Unsupervised Approaches
• Heatmap clustering
• SPADE
• viSNE
• Phenograph

See Table 1 of Diggins et al., Methods 2015 for list of unsupervised tools
Traditional Gating Overlooks Many Cells in Primary Samples

viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia.

El-ad David Amir\textsuperscript{1,2}, Kara L Davis\textsuperscript{2,3}, Michelle D Tadmor\textsuperscript{1,3}, Erin F Simonds\textsuperscript{2,3}, Jacob H Levine\textsuperscript{1,3}, Sean C Bendall\textsuperscript{1,3}, Daniel K Shenfeld\textsuperscript{1,3}, Smita Krishnaswamy\textsuperscript{1}, Garry P Nolan\textsuperscript{2,3,4} & Dana Pe'er\textsuperscript{1,4}

In all cases, the viSNE gate included cells that were not classified by the expert manually gated biaxial plots; these cells are labeled in gray in the viSNE map. Examination of the marker expression of these cells reveals that they are typically just beyond the threshold of one marker, but the viSNE classification is strongly supported based on the expression of all other markers. For example, in Figure 1d, wherein cells are colored for CD11b marker expression, the cells in the gated region express the canonical monocyte marker CD33 (Supplementary Fig. 1b). However, only 47% of these cells were classified as monocytes by the manual gating (Fig. 1b).

High-dimensional analysis of the murine myeloid cell system

Burkhard Becher\textsuperscript{1,4,5}, Andreas Schlitzer\textsuperscript{1,5}, Jinmiao Chen\textsuperscript{1,5}, Florian Mair\textsuperscript{2}, Hermi R Sumato\textsuperscript{1}, Karen Wei Weng Teng\textsuperscript{1}, Donovan Low\textsuperscript{1}, Christiane Ruedl\textsuperscript{1}, Paola Riccardi-Castagnoli\textsuperscript{1}, Michael Poidinger\textsuperscript{4}, Melanie Greter\textsuperscript{2}, Florent Ginhoux\textsuperscript{1} & Evan W Newell\textsuperscript{1}

Notably, whereas traditional biased gating strategies allowed for identification of only $54.7 \pm 2.6\%$ (mean ± s.e.m., $n = 3$ mice) of lung myeloid cells (different DC subsets, macrophages, monocytes, neutrophils), the automatic, computational approach identified nearly 100% of the cells ($96.6 \pm 1.0\%$ (mean ± s.e.m., $n = 3$ mice) accounted for by 14 predominant clusters).
## Major Steps in Most Single Cell Biology Workflows

<table>
<thead>
<tr>
<th>Data collection</th>
<th>1) Panel design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2) Data collection</td>
</tr>
<tr>
<td>Data processing</td>
<td>3) Cell event parsing</td>
</tr>
<tr>
<td></td>
<td>4) Scale transformation</td>
</tr>
<tr>
<td>Distinguishing initial populations</td>
<td>5) Live single cell gating</td>
</tr>
<tr>
<td></td>
<td>6) Focal population gating</td>
</tr>
<tr>
<td>Revealing cell subsets</td>
<td>7) Feature selection</td>
</tr>
<tr>
<td></td>
<td>8) Dimensionality reduction</td>
</tr>
<tr>
<td></td>
<td>9) Identify cell clusters</td>
</tr>
<tr>
<td></td>
<td>10) Cluster refinement</td>
</tr>
<tr>
<td>Characterizing cell subsets</td>
<td>11) Feature comparison</td>
</tr>
<tr>
<td></td>
<td>12) Model populations</td>
</tr>
<tr>
<td></td>
<td>13) Learn cell identity</td>
</tr>
<tr>
<td></td>
<td>14) Statistical testing</td>
</tr>
</tbody>
</table>

**How much can be automated?**

**Where do computers outperform humans?**

**How do we select tools and use them well?**

Diggins et al., *Methods* 2015
Teaching Computers To Spot Useful Patterns: Grouping Cells by Selected Features (e.g. Protein Expression)

1. HD cytometry!!
2. Woah, that’s a lot of data…
3. Computational tools
4. Biological knowledge

Adapted from LEGO 70803 Cloud Cuckoo Palace instructions 2014
Many Great Tools Exist, But Key Gaps Remain

A major gap in the field is in true learning of cell identity.
Key Analysis Concepts: Dimensionality Reduction, Transformation, Clustering, Modeling, Visualization, & Integration

**viSNE**
Amir et al.
*Nature biotech* 2013

**SPADE**
Qiu et al.
*Nature biotech* 2011

Diggins et al., *Methods* 2015
We Will Use Cytobank Software for viSNE & SPADE

Cytobank (www.cytobank.org) is a commercial tool for web-based data storage, annotation, analysis, and visualization.

30-day free trial with viSNE & SPADE: https://premier.cytobank.org/signup
Discussion Questions Covered in Today’s Course

1) What are key differences between tools (viSNE, SPADE, PCA)? What is the difference between transforming, clustering, and modeling data? What type of modeling are we doing (if any)?

2) What does non-linear vs. linear analysis mean? Does the data’s scale matter for analysis (arcsinh5, arcsinh15, linear)?

3) What do viSNE and SPADE settings do (viSNE iterations, SPADE downsampling & node #)? When should they be changed?

5) How does one compare new samples with a prior analysis? How do we test tools with expert gating?

6) What are some “red flags” indicating problems? What does a good viSNE or SPADE analysis run look like?
Onward, to the analysis!
Discovery and Characterization of Cell Subsets: Towards Machine Learning Cell Identity

1. Identify cell events
   - Blood cells, AML patient
   - Marrow cells, healthy donor

2. viSNE
   - Non-cancer cells in either sample
   - Leukemia blasts

3. SPADE
   - Phenotypically distinct cell subsets

4. Clustered heatmap
   - Discovered cell subsets
   - Protein signature
   - Clinical Features

Diggins et al., Methods 2015
### Single Cell Biology Workflow

| Data collection | 1) Panel design  
<table>
<thead>
<tr>
<th></th>
<th>2) Data collection</th>
</tr>
</thead>
</table>
| Data processing| 3) Cell event parsing  
|                | 4) Scale transformation |
| Distinguishing initial populations | 5) Live single cell gating  
|                | 6) Focal population gating |
| Revealing cell subsets | 7) Feature selection  
|                | 8) Dimensionality reduction  
|                | 9) Identify cell clusters  
|                | 10) Cluster refinement  
| Characterizing cell subsets | 11) Feature comparison  
|                | 12) Model populations  
|                | 13) Learn cell identity  
|                | 14) Statistical testing  |

---

Diggins et al., *Methods* 2015
Mass Cytometry: 35+ Dimensional Analytical Cytometry

Adapted from Bendall et al., Science 2011, Cytobank, & DVS Sciences
Mass Cytometry Data Pre-Processing

Data Acquisition (IMD -> FCS) → Concatenation → Normalization → Debarcoding → Transformation → Analysis

Resources:
- **Concatenation**: downloadable tool from Cytobank (http://support.cytobank.org/help/kb/cytobank-utilities/concatenating-fcs-files)
- **Barcoding**: Bodenmiller et al, Nature Biotechnology 2012 (http://www.nature.com/nbt/journal/v30/n9/full/nbt.2317.html)

Fink et al, Cytometry Part A 2013

FCS File 1 + FCS File 2 + FCS File 3

arcsinh(x/15)
Steps 1-3: Getting Started & Preparing for viSNE

Workflow summary:
1) **Clone** experiment “Diggins et al., Methods 2015 – Step 1”
2) Set the **Scales** so that the scale argument is 15 for all channels.
3) **Gate** for “Intact cells”

- Navigate to Diggins et al., Methods 2015 – Step 1
- Click the experiment name to load it

https://premium.cytobank.org/cytobank/experiments/44410/
Steps 1-3: Getting Started & Preparing for viSNE

Workflow summary:
1) **Clone** experiment “Diggins et al., Methods 2015 – Step 1”
2) Set the **Scales** so that the scale argument is 15 for all channels.
3) **Gate** for “Intact cells”

- Make a clone using Actions => Cloning => Selective Clone
- Edit the name, uncheck all except “clone attachments and protocols”, press Clone button.

https://premium.cytobank.org/cytobank/experiments/44410/
Have you ever noticed two peaks within the cells that are biologically 100% negative for a marker?

Results from bad scaling (poor transformation) and it can be an issue for computational analysis.

Scaling is important in both mass and fluorescence cytometry.

Scaling Matters for Measuring Distance

For fluorescent flow cytometry data a biexponential or arcsinh transformation corrects the scale near zero.

Since computational analysis techniques compare distance similar to what a person does when looking at a plot, these techniques can identify artificial populations near zero (see C and D) if data are not appropriately transformed prior to analysis.

More information: https://my.vanderbilt.edu/irishlab/protocols/scales-and-transformation/
Inappropriate Scaling Can Lead to False Population Discovery

Here, CD8 was under-scaled so that an artificial ‘hole’ in the graph existed around zero. This created the false impression of two CD8 populations in this sample gated as CD8 negative. SPADE population discovery treated this as significant.
Steps 1-3: Getting Started & Preparing for viSNE

Workflow summary:
1) **Clone** experiment “Diggins et al., Methods 2015 – Step 1”
2) Set the **Scales** so that the scale argument is 15 for all channels.
3) **Gate** for “Intact cells”

- Open the Scales and set the scale argument to 15 for all channels, then press Apply.
- Choose “OK” when the popup asks if you really want to do it.

https://premium.cytobank.org/cytobank/experiments/44411/
Steps 1-3: Getting Started & Preparing for viSNE

Workflow summary:
1) **Clone** experiment “Diggins et al., Methods 2015 – Step 1”
2) Set the **Scales** so that the scale argument is 15 for all channels.
3) **Gate** for “Intact cells”

- Go into **Gating**, change the y-axis to “Cell Length” (Event Length) and the x-axis to NA191 (Intercalator).

- Draw a polygon gate like the one below and call it “Intact cells”.

- **Apply** the gate.

https://premium.cytobank.org/cytobank/experiments/44411/
Steps 1-3: Getting Started & Preparing for viSNE

Workflow summary:

1) **Clone** experiment “Diggins et al., Methods 2015 – Step 1”
2) Set the **Scales** so that the scale argument is 15 for all channels.
3) **Gate** for “Intact cells”

- Click on the “Intact cells” gate in the list of gates to the right.
- Make the gate a tailored gate by clicking the “Tailored” radio circle.
- Select the other file (Step1b_normal_marrow)
- Move the gate’s vertices to include 100% of normal marrow events.
- Make sure to **Apply** the gate. Optional: “Check gate”.

Note: we’re doing this because intercalator levels tend to vary from run to run (in contrast with antibody staining).

https://premium.cytobank.org/cytobank/experiments/44411/
Discovery and Characterization of Cell Subsets: Towards Machine Learning Cell Identity

Diggins et al., Methods 2015
## Single Cell Biology Workflow

| Data collection          | 1) Panel design  
<table>
<thead>
<tr>
<th></th>
<th>2) Data collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data processing</td>
<td>3) Cell event parsing</td>
</tr>
<tr>
<td></td>
<td>4) Scale transformation</td>
</tr>
<tr>
<td>Distinguishing initial populations</td>
<td>5) Live single cell gating</td>
</tr>
<tr>
<td></td>
<td>6) Focal population gating</td>
</tr>
<tr>
<td>Revealing cell subsets</td>
<td>7) Feature selection</td>
</tr>
<tr>
<td></td>
<td>8) Dimensionality reduction</td>
</tr>
<tr>
<td></td>
<td>9) Identify cell clusters</td>
</tr>
<tr>
<td></td>
<td>10) Cluster refinement</td>
</tr>
<tr>
<td>Characterizing cell subsets</td>
<td>11) Feature comparison</td>
</tr>
<tr>
<td></td>
<td>12) Model populations</td>
</tr>
<tr>
<td></td>
<td>13) Learn cell identity</td>
</tr>
<tr>
<td></td>
<td>14) Statistical testing</td>
</tr>
</tbody>
</table>

---

**Expert gating**

- viSNE + expert gating

---

Diggins et al., *Methods* 2015
Current Goal: Get to Figure 1 from Diggins et al. with viSNE
Stochastic Neighbor Embedding (SNE)

- SNE used for image recognition
- 60,000 handwritten greyscale images
- 28x28 pixels each

Example: 6x6 Pixel Image

Vectorize (1x36)

viSNE / t-SNE Arranges Cells in 2D by Multi-D Similarity

Healthy human blood, mass cytometry, 26 markers measured, viSNE analysis tool

Animation created by Cytobank team from iterations of viSNE / t-SNE using PBMC (26 features)
viSNE / t-SNE Arranges Cells in 2D by Multi-D Similarity

Healthy human blood, mass cytometry, 26D viSNE analysis

Animation created by Cytobank team from iterations of viSNE / t-SNE using PBMC (26 features)
Viewing Expert Gates with viSNE Reveals Cyto Incognito

Healthy human blood, mass cytometry, 26D viSNE analysis

Created with Chris Ciccolella, Cytobank
Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

Healthy human blood, mass cytometry, 26D viSNE analysis

Created with Chris Ciccolella, Cytobank
Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

Healthy human blood, mass cytometry, 26D viSNE analysis

Created with Chris Ciccolella, Cytobank

\[ i = 45 \]
Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

Healthy human blood, mass cytometry, 26D viSNE analysis

Created with Chris Ciccolella, Cytobank
Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

Healthy human blood, mass cytometry, 26D viSNE analysis

Created with Chris Ciccolella, Cytobank
Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

Healthy human blood, mass cytometry, 26D viSNE analysis

Created with Chris Ciccorella, Cytobank
Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

Healthy human blood, mass cytometry, 26D viSNE analysis

Created with Chris Ciccolella, Cytobank
Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

Healthy human blood, mass cytometry, 26D viSNE analysis

Created with Chris Ciccorella, Cytobank
Viewing Expert Gates with viSNE Reveals *Cyto Incognito*

Healthy human blood, mass cytometry, 26D viSNE analysis

Created with Chris Ciccorella, Cytobank
Viewing Expert Gates with viSNE Reveals Cyto Incognito

Healthy human blood, mass cytometry, 26D viSNE analysis

Cyto incognito (Cells overlooked or hidden in expert gating)

Created with Chris Ciccolella, Cytobank
Phenograph Adds Fast Clustering & Meta-Analysis to viSNE

Levine et al., Cell 2015
Step 4: Run viSNE on Intact Cells from Both Samples

Workflow summary:

4) Run viSNE.

- Pick viSNE => New viSNE Analysis
- Name it “viSNE of AML and healthy marrow”

https://premium.cytobank.org/cytobank/experiments/44411/
Step 4: Run viSNE on Intact Cells from Both Samples

Workflow summary:

4) Run **viSNE**.

- Pick **Intact Cells** from the Population selection area.
- Select both files to include in the viSNE analysis.

---

https://premium.cytobank.org/cytobank/experiments/44411/
Step 4: Run viSNE on Intact Cells from Both Samples

Workflow summary:

4) Run viSNE.

- Choose Equal subsampling and leave the “events per population” at the default of 50,000 (so 100,000 total events).

https://premium.cytobank.org/cytobank/experiments/44411/
Step 4: Run viSNE on Intact Cells from Both Samples

Workflow summary:

4) Run viSNE.

- Scroll down, and in the channel selector, pick the channels you want to be used in the comparison analysis (to “make the map”).
- Pick “all 27 markers”. Search on “CD” and “select all” (selects 26), then clear the selection and add HLA-DR to the list. Click “Done” and “Run viSNE Analysis.”

https://premium.cytobank.org/cytobank/experiments/44411/
PBMC, AML Patient, Day 0 (pre-treatment)

viSNE

Populations gated

Non-blasts, AML blood
AML blasts

CD8+ T cells
NKs
B cells
CD4+ T cells

Population interpretations
AML subsets

Next: View All Channels to Make Figure 1 from Diggins et al.

Diggins, et al Methods 2015
While viSNE runs…

A brief interlude about topology
Is There ‘Inherent Topology’ in the Data?

An illustration of PCA. **a)** A data set given as 3-dimensional points. **b)** The three orthogonal Principal Components (PCs) for the data, ordered by variance. **c)** The projection of the data set into the first two PCs, discarding the third one.

Effects of dimensionality reduction on an inherently non-linear data set. **a)** The original data given as a two-dimensional set. **b)** PCA identifies two PCs as contributing significantly to explain the data variance. **c)** However, the inherent topology (connectivity) of the data helps identify the set as being one-dimensional, but non-linear.
Mapping Topology Matters When Considering “Travel”
Isomap at work on the "swiss roll" data set. **a)** The input data are given as three-dimensional, but are really two-dimensional in nature. **b)** Each point in the data set is connected to its neighbors to form a neighborhood graph, overlaid in light grey. Geodesic distances are approximated by computing **shortest paths along the neighborhood graph** (red). **c)** MDS applied to the geodesic distances has the effect of "unrolling" the swiss roll into its natural, two-dimensional parameterization. The neighborhood graph has been overlaid for comparison. Now, the Euclidean distances (in blue) approximate the original geodesic distances (in red).
**Geodesic Distance Measurements**

**May Better Capture Developmental Progressions**

*Lydia E. Kavraki, Geometric Methods in Structural Computational Biology*

**Geodesic distance.** The geodesic distance between the two red points is the length of the geodesic path, which is the shortest path between the points, that lies on the surface.

“**Creode** is a neologism coined by the biologist C.H. Waddington to represent the developmental pathway followed by a cell as it grows to form part of a specialized organ. Combining the Greek roots for "necessary" and "path," the term was inspired by the property of regulation.”

http://en.wikipedia.org/wiki/Creode
Geodesic Distance Measurements May Better Capture Developmental Progressions

Lydia E. Kavraki, *Geometric Methods in Structural Computational Biology*

**Geodesic distance.** The geodesic distance between the two red points is the length of the geodesic path, which is the shortest path between the points, *that lies on the surface.*

OK, back to work…
Steps 5-7: Gating on viSNE & Exporting for SPADE

Workflow summary:

5) Gate on the t-SNE axes for “AML PB blasts” (CD45\text{lo}) and “AML PB non-blasts” (CD45\text{hi})
6) Export gated populations as 4 new FCS files & discard the empty one.
7) Concatenate the 3 remaining files into a merged file.

- Change the z-axis coloring channel to CD45.
- Using the knowledge the AML blasts are CD45\text{lo}, draw gates for “AML blast cells” and “non-blast cells”.
- Use “Check Gates” button to make sure gates work well on both files. Healthy marrow should have <3% cells in the AML blast gate.
- Apply & Return.

https://premium.cytobank.org/cytobank/experiments/44415/
Steps 5-7: Gating on viSNE & Exporting for SPADE

Workflow summary:

5) Gate on the t-SNE axes for “AML PB blasts” (CD45$^{lo}$) and “AML PB non-blasts” (CD45$^{hi}$)

6) Export gated populations as 4 new FCS files & discard the empty one.

7) Concatenate the 3 remaining files into a merged file.

- Set up an illustration to export the viSNE gated data: Pick 2 Populations (“AML blast cells” & “Non-blast cells”). Pick one Channel (CD45). Activate “FCS files”. Update & Save.
Steps 5-7: Gating on viSNE & Exporting for SPADE

Workflow summary:

5) Gate on the t-SNE axes for “AML PB blasts” (CD45\text{lo}) and “AML PB non-blasts” (CD45\text{hi})

6) Export gated populations as 4 new FCS files & discard the empty one.

7) Concatenate the 3 remaining files into a merged file.

• Export viSNE gated data: Actions => Cloning => Split Files by Population

https://premium.cytobank.org/cytobank/experiments/44415/
Steps 5-7: Gating on viSNE & Exporting for SPADE

Workflow summary:

5) **Gate** on the t-SNE axes for “AML PB blasts” (CD45<sup>lo</sup>) and “AML PB non-blasts” (CD45<sup>hi</sup>)

6) Export gated populations as 4 new FCS files & discard the empty one.

7) **Concatenate** the 3 remaining files into a merged file.

- Name the experiment “Diggins et al. - Export for SPADE” and click “Create Experiment”

https://premium.cytobank.org/cytobank/experiments/44415/
Steps 5-7: Gating on viSNE & Exporting for SPADE

Workflow summary:

5) **Gate** on the t-SNE axes for “AML PB blasts” (CD45$^{lo}$) and “AML PB non-blasts” (CD45$^{hi}$)

6) Export gated populations as 4 new FCS files & discard the empty one.

7) **Concatenate** the 3 remaining files into a merged file.

- Download the files and concatenate (merge) into 1 file.

http://blog.cytobank.org/2013/10/29/new-fcs-file-concatenation-tool/

https://premium.cytobank.org/cytobank/experiments/44415/
Discovery and Characterization of Cell Subsets: Towards Machine Learning Cell Identity

Diggins et al., Methods 2015
## Single Cell Biology Data Analysis

| Data collection | 1) Panel design  
| 2) Data collection |
|------------------|-----------------|
| Data processing  | 3) Cell event parsing  
| 4) Scale transformation |
| Distinguishing initial populations | 5) Live single cell gating  
| 6) Focal population gating |
| Revealing cell subsets | 7) Feature selection  
| 8) Dimensionality reduction  
| 9) Identify cell clusters  
| 10) Cluster refinement |
| Characterizing cell subsets | 11) Feature comparison  
| 12) Model populations  
| 13) Learn cell identity  
| 14) Statistical testing |

**Feature selection**

**SPADE: cell subset clustering**

- AML blasts
- Non-blasts, in AML
- Healthy marrow

- AML blasts
- Non-blasts, in AML
- Healthy marrow

Diggins et al., *Methods* 2015
Current Goal: Make Figure 2 from Diggins et al. with SPADE

Diggins, et al Methods 2015
SPADE Extracts Population Hierarchies from Multi-D Space

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

Peng Qiu1,2, Erin F Simonds3, Sean C Bendall1, Kenneth D Gibbs Jr3, Robert V Bruggner3, Michael D Linderman4, Karen Sachs3, Garry P Nolan3 & Sylvia K Plevritis1

VOLUME 29 NUMBER 10 OCTOBER 2011 NATURE BIOTECHNOLOGY

SPADE method:

1) Get Data
2) Downsampling (preserves rare subsets)
3) Cluster (group by similarity)
4) Project in 2D (minimum spanning tree)

SPADE stands for ‘Spanning-tree Progression Analysis of Density-normalized Events'. A graphical interface and updates for fluorescence datasets are implemented on Cytobank.

Irish Lab @ Vanderbilt
SPADE Trees Depict Multidimensional Similarity
(not necessarily developmental relationships)

SPADE trees can sometimes be reorganized so that closely related branches appear far apart. In order to remove loops from the SPADE tree, breakpoints may be added in unexpected places.

Irish Lab @ Vanderbilt
SPADE trees can sometimes be reorganized so that closely related branches appear far apart. In order to remove loops from the SPADE tree, breakpoints may be added in unexpected places.
SPADE Trees Depict Multidimensional Similarity
(not necessarily developmental relationships)

SPADE trees can sometimes be reorganized so that closely related branches appear far apart. In order to remove loops from the SPADE tree, breakpoints may be added in unexpected places.

Irish Lab @ Vanderbilt
SPADE trees can sometimes be reorganized so that closely related branches appear far apart. In order to remove loops from the SPADE tree, breakpoints may be added in unexpected places.

Irish Lab @ Vanderbilt
SPADE trees can sometimes be reorganized so that closely related branches appear far apart. In order to remove loops from the SPADE tree, breakpoints may be added in unexpected places.

Irish Lab @ Vanderbilt
SPADE trees can sometimes be reorganized so that closely related branches appear far apart. In order to remove loops from the SPADE tree, breakpoints may be added in unexpected places.
Steps 8 and 9: SPADE Clustering on t-SNE Channels

Workflow summary:

1) **Clone** experiment “Diggins et al., Methods 2015 – Step 8”
2) Set the **Scales** so that the scale argument is 15 for all channels.
3) Run **SPADE**.

- Click the experiment name to load it.
- Clone the experiment, as before. Actions => Cloning => Selective Clone.
- Click **Scales** to edit the scales and set the argument to 15 for all channels. Click Apply and then close this window.
- The cells have already been gated, so no gating is needed.
- Choose **SPADE** => New SPADE Analysis & give it a name.

https://premium.cytobank.org/cytobank/experiments/44412
Steps 8 and 9: SPADE Clustering on t-SNE Channels

Workflow summary:
1) **Clone** experiment “Diggins et al., Methods 2015 – Step 8”
2) Set the **Scales** so that the scale argument is 15 for all channels.
3) Run **SPADE**.

- Set SPADE to use 100 nodes & 1% downsampling.
- Select just the 2 t-SNE channels.
- Click the green bar to run the analysis.

https://premium.cytobank.org/cytobank/experiments/44412
Steps 8 and 9: SPADE Clustering on t-SNE Channels

Workflow summary:
1) **Clone** experiment “Diggins et al., Methods 2015 – Step 8”
2) Set the **Scales** so that the scale argument is 15 for all channels.
3) Run **SPADE**.

- Explore the data in the SPADE viewer.
- Make your node size larger. Set the x/y axes to CD45 and CD34. Set the Node Color Parameter to CD34.
- See if you can find the AML blasts and bubble them.
- View various Parameters.

https://premium.cytobank.org/cytobank/experiments/44413/spade/12274
Steps 8 and 9: SPADE Clustering on t-SNE Channels

Workflow summary:
1) **Clone** experiment “Diggins et al., Methods 2015 – Step 8”
2) Set the **Scales** so that the scale argument is 15 for all channels.
3) Run **SPADE**.

- Scroll way down and click “Select All Nodes”.
- In the “Actions” panel, Choose “Download Statistics Table”
- In the ZIP file, you will find a folder called “By Sample” that has spreadsheets of median expression by node for all samples. These can be sorted, filtered, and clustered.

https://premium.cytobank.org/cytobank/experiments/44413/spade/12274
Discovery and Characterization of Cell Subsets: Towards Machine Learning Cell Identity

1. Identify cell events using mass cytometry.
   - Blood cells, AML patient
   - Marrow cells, healthy donor

2. viSNE for non-cancer cells in either sample.
   - Leukemia blasts

3. SPADE for phenotypically distinct cell subsets.

4. Clustered heatmap for discovered cell subsets.
   - Protein signature
   - Clinical Features

Diggins et al., *Methods* 2015
## Single Cell Biology Data Analysis

| Data collection | 1) Panel design  
| Data processing | 2) Data collection  
| Distinguishing initial populations | 3) Cell event parsing  
| Revealing cell subsets | 4) Scale transformation  
| Characterizing cell subsets | 5) Live single cell gating  
| | 6) Focal population gating  
| | 7) Feature selection  
| | 8) Dimensionality reduction  
| | 9) Identify cell clusters  
| | 10) Cluster refinement  
| | 11) Feature comparison  
| | 12) Model populations  
| | 13) Learn cell identity  
| | 14) Statistical testing  

---

**Grouped populations of cells**

**Heatmap: clustering populations**

**Compare and visualize populations across datasets**

Diggins et al., *Methods* 2015
Finally, The Fun/Challenging Part: “Call” Subset Identity and Characterize Subset Features

3. Cluster SPADE Nodes and Display as Heatmap of Medians

Note: Heatmaps allow comparison of data from different viSNE and SPADE runs, different tools, different machines, etc.
Finally, The Fun/Challenging Part: “Call” Subset Identity and Characterize Subset Features

3. Cluster SPADE Nodes and Display as Heatmap of Medians

Note: Heatmaps allow comparison of data from different viSNE and SPADE runs, different tools, different machines, etc.
Finally, The Fun/Challenging Part: “Call” Subset Identity and Characterize Subset Features

3. Cluster SPADE Nodes and Display as Heatmap of Medians

Note: Heatmaps allow comparison of data from different viSNE and SPADE runs, different tools, different machines, etc.

Finally, The Fun/Challenging Part: “Call” Subset Identity and Characterize Subset Features

3. Cluster SPADE Nodes and Display as Heatmap of Medians

Note: Heatmaps allow comparison of data from different viSNE and SPADE runs, different tools, different machines, etc.
Heatmaps Also Visualize Other Data Types (e.g. Stratified Clinical Outcomes) and Compare Across Analysis Runs

Heatmaps visualize across integrated data types, e.g. clinical outcomes, cytogenetics, & signaling profiles

Irish, et al Cell 2004
Other analysis tools
Principal Component Analysis

Feature A

Feature B

Feature C

Principal Component 1

$X_{A1} A + X_{B1} B + X_{C1} C$

Principal Component 2

$X_{A2} A + X_{B2} B + X_{C2} C$

Principal Component 3

$X_{A3} A + X_{B3} B + X_{C3} C$

PCA used to Reduce Dimensionality of CyTOF Data

Newell et al 2012, *Immunity*
Mixture Modeling

SWIFT

A: Initial sub-populations:
May be skewed;
May overlap;
May have a high dynamic range.

1. EM fitting: The EM algorithm fits data to a specified number of Gaussians, by weighted, iterative sampling. Large asymmetric peaks may be split, but rare peaks may not separate.

B: 1+2

2. Splitting: Each cluster from Step 1 is tested by LDA for multiple modes in all combinations of dimensions. Clusters are split if necessary (using EM), until all are unimodal.

3. Merging: All cluster pairs are tested, and merged if the resulting cluster is unimodal in all dimensions. Agglomerative merging prevents over-merging due to ‘bridging’ Gaussians.

FLAME

Mosmann et al, 2014 Cytometry A

Pyne et al, 2009 PNAS
Automated Clustering and Population Identification Methods Based on Density

SamSpectral

FLOCK

Zare et al, 2010 BMC Bioinformatics

Qian et al, 2010 Cytometry B Clin Cytom
Wanderlust Identifies Phenotypic Progression

Bendall, *Cell* 2014
ISOMAP guided analysis

to compare overall phenotypic relatedness of populations of neutrophil-like and eosinophil-like cells. Top, cells color-coded by DensVM cluster number are plotted by their scores for ISOMAP dimensions 1 and 2. Binned median expression of defining markers (middle) and the tissue composition (percentage of each cluster as a fraction of total granulocytes from each tissue, bottom) of cells along this phenotypic progression defined by ISOMAP dimension 1 and DensVM clusters 1–7 are plotted.
Citrus: Supervised Population Finding

Automated identification of stratifying signatures in cellular subpopulations

Robert V. Bruggner\textsuperscript{a,b}, Bernd Bodenmiller\textsuperscript{a}, David L. Dill\textsuperscript{c}, Robert J. Tibshirani\textsuperscript{d,f,1}, and Garry P. Nolan\textsuperscript{b,1}

\textsuperscript{a}Biomedical Informatics Training Program, Stanford University Medical School, Stanford, CA 94305; \textsuperscript{b}Baxter Laboratory for Stem Cell Biology, Department of Microbiology and Immunology, and Departments of \textsuperscript{c}Computer Science, \textsuperscript{d}Health Research and Policy, and \textsuperscript{f}Statistics, Stanford University, Stanford, CA 94305; \textsuperscript{1}Institute of Molecular Life Sciences, University of Zurich, CH-8097 Zurich, Switzerland

Contributed by Robert J. Tibshirani, May 14, 2014 (sent for review February 12, 2014)

\begin{itemize}
\item[(i)] Patient 1
  \begin{itemize}
  \item Cell 1
  \item Cell 2
  \item \ldots
  \item Cell \textit{n}
  \end{itemize}
\item[(ii)] Marker 2
\item[(iii)] Marker 1
\item[(iv)] Patient Status
  \begin{itemize}
  \item Healthy
  \item Diseased
  \item \ldots
  \item Diseased
  \end{itemize}
\item[(v)] Regularized Regression Model
\item[(vi)] Subset A abundance
\item[(vii)] Subset A
\end{itemize}

\textsuperscript{a}Time Dependent ROC Curves

\begin{itemize}
\item Method
\item \textit{a} citrus
\item \textit{b} flowType
\item \textit{c} dataset
\item \textit{d} test
\item \textit{e} train
\end{itemize}

\begin{itemize}
\item citrus/training AUC = 0.69 (95\% CI 0.58–0.81)
\item citrus/testing AUC = 0.80 (95\% CI 0.70–0.92)
\item flowType/training AUC = 0.65 (95\% CI 0.52–0.79)
\item flowType/testing AUC = 0.52 (95\% CI 0.36–0.69)
\end{itemize}

Bruggner et al., PNAS 2014
Final notes & conclusions...
Computational Tools + HD Cytomics Together Are Powering A New Era in Clinical Oncology & Immunity

1) Need pre-treatment prognosis & prediction
   If diagnostic scheme does not provide actionable information, fix it.
   Who will benefit from expensive cell based therapies (~$500,000 ea.)?
   In the absence of mutations, clinical response can be predicted by cell profiling.

2) Need to monitor treatment longitudinally
   See early whether patient responded / adjust treatment, as needed.
   Monitor whether treatment is still required.

3) Need to check multiple biomarkers with one test
   As with genetic tests, multiplexing biomarkers will give more information per sample, catch the unexpected, and cost less than repeated testing.

4) Need to monitor biomarkers on all cell types
   PD-L1 is a great example – expressed by many cell types & can be activated.

5) Need to characterize all cell types to monitor cancer
   Evolving cancer cells adopt unexpected phenotypes.
Conclusions: Data Analysis & Mapping Cell Identity

Workflow summary:
1) viSNE with minimal pre-gating
2) SPADE, works especially well on t-SNE axes
3) Heatmap to compare with other data, do statistical tests

1) A modular workflow allows comparison of different tools at each step. Synthetic channels are useful (e.g. t-SNEs).

2) Non-linear transformation may help; tool selection depends on ‘what works’, biology, & data shape.

3) Many outstanding tools are available (see Diggins et al. for reference list). Still need tools that learn.
Discussion Questions Covered in Today’s Course

1) What are key differences between tools (viSNE, SPADE, PCA)? What is the difference between transforming, clustering, and modeling data? What type of modeling are we doing (if any)?

2) What does non-linear vs. linear analysis mean? Does the data’s scale matter for analysis (arcsinh5, arcsinh15, linear)?

3) What do viSNE and SPADE settings do (viSNE iterations, SPADE downsampling & node #)? When should they be changed?

5) How does one compare new samples with a prior analysis? How do we test tools with expert gating?

6) What are some “red flags” indicating problems? What does a good viSNE or SPADE analysis run look like?
Acknowledgements & Thank you!

L to R: Deon Doxie, Mikael Roussel, Kirsten Diggins, Cara Wogsland, Jonathan Irish, Brent Ferrell, Nalin Leelatian, Hannah Polikowsky, Allison Greenplate

Oslo Collaborators
June Myklebust
Kanutte Huse

VU Collaborators
Madan Jagasia
Michael Savona
Jeff Sosman
Doug Johnson
Pierre Massion
Rebecca Ihrie
Ann Richmond

NIH/NCI
R00 CA143231 (Irish Lab)
F31 CA199993 (Greenplate)
T32 CA009592 (Doxie)
R25 CA136440 (Diggins)
F31 CA203383 (Diggins)
K12 CA090625 (Ferrell)
VISP (Leelatian)

Vanderbilt-Ingram Cancer Center (VICC)
Ambassadors, Hematology Helping Hands

Vanderbilt University Discovery

Norwegian Cancer Society (Huse)
Swiss Foundation Nuovo-Soldati (Roussel)
R/BIOCONDUCTOR
R Flow Cytometry Data Analysis

Arcsinh Transformation of Data

Visualization

Gate T Cells on CD4 and CD8 Expression

Compare T cell Subset Frequencies Between Individuals
What is R/Bioconductor?

- **R**: statistical and graphical programming language/environment
- **Bioconductor**: provides open source packages for bioinformatics analysis in R
Anatomy of an FCS File in R

**FlowSet (set of FCS files)**
- FlowFrame1 (fcs file)
- FlowFrame2 (fcs file)
- FlowFrame3 (fcs file)
  ...

**FlowFrame (fcs file)**

**Exprs**
- Fluorescence intensity matrix

<table>
<thead>
<tr>
<th>Cell</th>
<th>Parameter1</th>
<th>Parameter2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MFI_1-1</td>
<td>MFI_1-2</td>
</tr>
<tr>
<td>2</td>
<td>MFI_2-1</td>
<td>MFI_2-2</td>
</tr>
</tbody>
</table>

```r
> exprs(FlowFrame)
```

**Parameters**
- Annotated data frame
- Info about parameters (stains)

- rowNames: $P1, P2, P3...
- varLabels: Name, desc...max range
- varMetadata: labelDescription

```r
> parameters(FlowFrame)
```

**Description**
- List of information from instrument

- $FCSversion: [1] "3"
- $BEGINANALYSIS: [1] "0"
- $ENDANALYSIS: [1] "0"
- ...

```r
> description(FlowFrame)
```
Flow Analysis in R

- **flowCore**: access data in fcs files
- **flowQ**: quality control for flow data
- **flowStats**: statistical analysis of flow data
- **flowViz**: visualization of flow data
- **flowClust**: mixture-modeling to find clusters
- **flowMerge**: merge clusters from mixture modeling
- **flowUtils**: parse Gating-ML files
- **flowFP**: transforms high dimensional data using modeling to facilitate data analysis
R package flowCore
- Read fcs files into R and perform basic analysis functions
- View and access intensity matrix

```r
> head(exprs(TcellData$Spec001a))
FSC-A  FSC-H  FSC-W  SSC-A  SSC-H  SSC-W  FITC-A  PerCP-Cy5-5-A  APC-A  Alexa Fluor 700-A  APC-H7-A  V450-A
[1,]  74313.46  64310  75730.17  26133.92  28082  61036.37  345.28   1647.36   1310.4  688.5  100.8  3174.22
[2,]  54602.26  45362  78885.72  25446.72  16436  101464.84  236.08  2462.72  2231.1  1161.9  846.9  8924.06
[3,]  67478.18  49500  89338.38  15140.32  17640  56249.21  143.52  2442.96  2241.0  1305.0  687.6  7780.16
[4,]  83260.04  64328  84623.56  31904.08  21015  99493.97  295.36  2546.96  2165.4  1104.3  538.2  6496.02
[5,]  54679.66  43581  82225.88  14788.00  15760  61410.89  260.00  1998.88  2108.7  945.9  346.5  9673.54
[6,] 107701.24  82691  84337.73  30740.32  23988  83983.55  101.92  5040.88  3944.7  2043.0  1160.1  5296.38
V500-A  PE-A  PE-Cy7-A  Time
[1,]  6639.540  66.33  102.96  0.4
[2,]  7280.280 1548.36 10053.45  0.7
[3,]  7010.180 1472.13  9365.40  2.1
[4,]  7343.300 1545.39  5906.34  2.1
[5,]  9341.430 1074.15  3357.09  3.0
[6,] 1669.520 1062.27 16261.74  3.2
>
> print(TcellData$Spec001a)
flowFrame object 'Spec001a'
with 32423 cells and 16 observables:

$name      $desc      $range    $minRange  $maxRange
$p1  FSC-A    FSC-A   262144     0     262144
$p2  FSC-H    FSC-H   262144     0     262144
$p3  FSC-W    FSC-W   262144     0     262144
$p4  SSC-A    SSC-A   262144     0     262144
$p5  SSC-H    SSC-H   262144     0     262144
$p6  SSC-W    SSC-W   262144     0     262144
$p7  FITC-A   CD57    262144    -111    262144
$p8  PerCP-Cy5-5-A CD28  262144    -111    262144
$p9  APC-A    CD27  262144    -111    262144
$p10 Alexa Fluor 700-A CD56  262144    -111    262144
$p11 APC-H7-A   CD8   262144    -111    262144
$p12 V450-A    CD3   262144    -111    262144
$p13 V500-A    CD4   262144    -111    262144
$p14 PE-A    CCR7   262144    -111    262144
$p15 PE-Cy7-A CD45RA  262144    -111    262144
$p16 Time    Time   262144    -111    262144

250 keywords are stored in the 'description' slot
```
flowViz

- Contour plot, density plot, scatter plot, trellis plot, and histograms for flow data

Density Dot (flowPlot(…))

Contour Plot (contour(…))

Histograms (densityplot(…))
flowQ

- Compare number of cells between files
- Check for margin events
- Check for time anomalies
- Data normalization
- KL Divergence
flowStats

- Functions for statistical analysis of flow data
  - Probability binning + Chi-squared test
  - Create filters for high density regions (auto-gating)
  - Quadrant gating
  - Methods for data normalization