

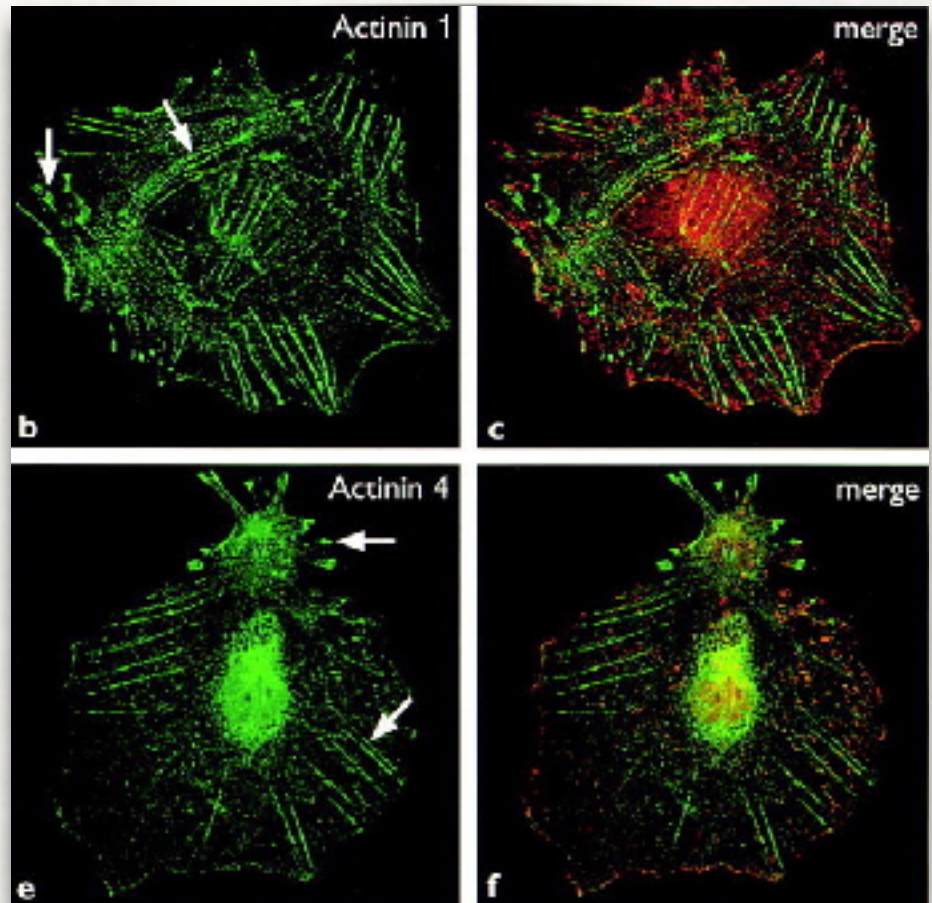
# AUTOMATED LEARNING OF SUBCELLULAR VARIATION AMONG PUNCTUATE PATTERNS AND A GENERATIVE MODEL OF THEIR RELATION TO MICROTUBULES

---

Gregory R. Johnson, Jieyue Li, Aabid Shariff, Gustavo K. Rohde, Robert F. Murphy

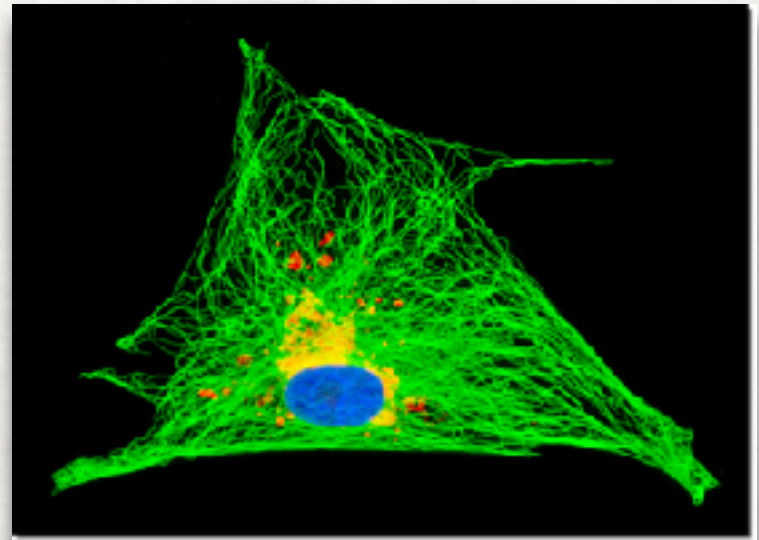
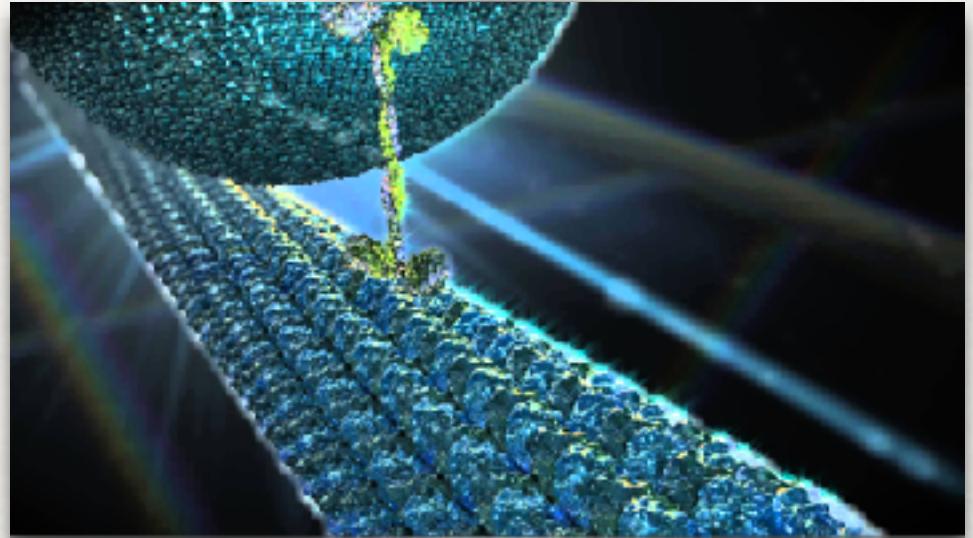
# PUNCTATE PROTEIN PATTERNS & FLUORESCENCE

- Give Rise to Sub-Cellular Spatial Protein Distributions
- **High Specificity** with high temporal and spatial resolution of living cells
- Protein localization and compartmentalization central to functionality
  - **Protein Conformations:** Compartments have varying chemical and physical characteristics influencing
  - **Metabolic Activity:** Organelles are locations of specialized functions in the cell



# MICROTUBULES

- “Highway” of the cells
- Filamentous intracellular **Structural Components**
- Part of the **Cytoskeleton**
- Readily identifiable when fluoresced
- Involved in
  - Nucleic and Cell Division
  - Cell Structure
  - Intracellular Protein Transport





# THE PROBLEM

## QUANTIFYING PATTERNS

- Established methods **Unable to Recognize** certain sub-patterns of major organelle types
  - Complex interconnectivity between proteins and underlying cell structures
  - **Variation Between Different Cell Types** makes pattern generalizations even more difficult
  - **Distinguishing at High Resolution** (membrane-bound organelles vs. macromolecular complexes) so far inconclusive
- **Quantification of Spatiotemporal Patterns** beyond human interpretation needed for further work in cell biochemistry and behavior simulations
- **Generation of Patterns** for incomplete pattern families or novel, yet similar proteins for simulation purposes

# THE PAPER'S AIM

- Fluorescence microscopy images of cells from A-431, U-2OS and U-251MG cell lines from Human Protein Atlas
- **Label all Images of proteins With Unclear Subcellular Pattern** annotations ("vesicles" or "cytoplasm")
- Already described systems for building image-derived, 2D or 3D generative models of distributions of other punctate organelles or microtubules within cells in previous papers
  - **Model Microtubule-Puncta Relationship** not present in previous model to enhance pattern recognition
- **Create Generative Model of Sub-Cellular Patterns**

# THE METHOD

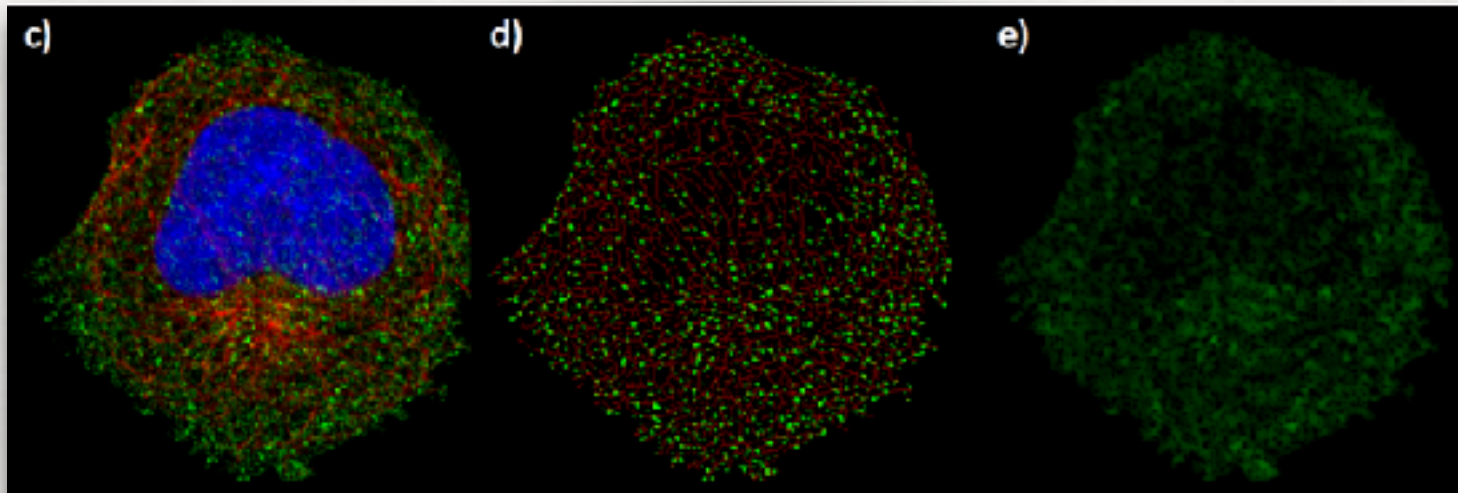
## THE FOUNDER BASELINE

- **11 Founder Proteins**
  - Subcellar location reasonably **Well Characterized**
  - Found in **11 Specific, Distinguishable** types of punctate patterns
  - Showed **Similar Pattern Across** all three cell types
  - Represent **Wide Range of** membrane and non-membrane bound **Compartments**
- Calculated feature matrices for all cells for each combination of **Eleven Proteins X Three Cell Lines**
- **Verification** of Relevance through Inspection & Principal Component Analysis

# THE METHOD

## IMAGE PROCESSING

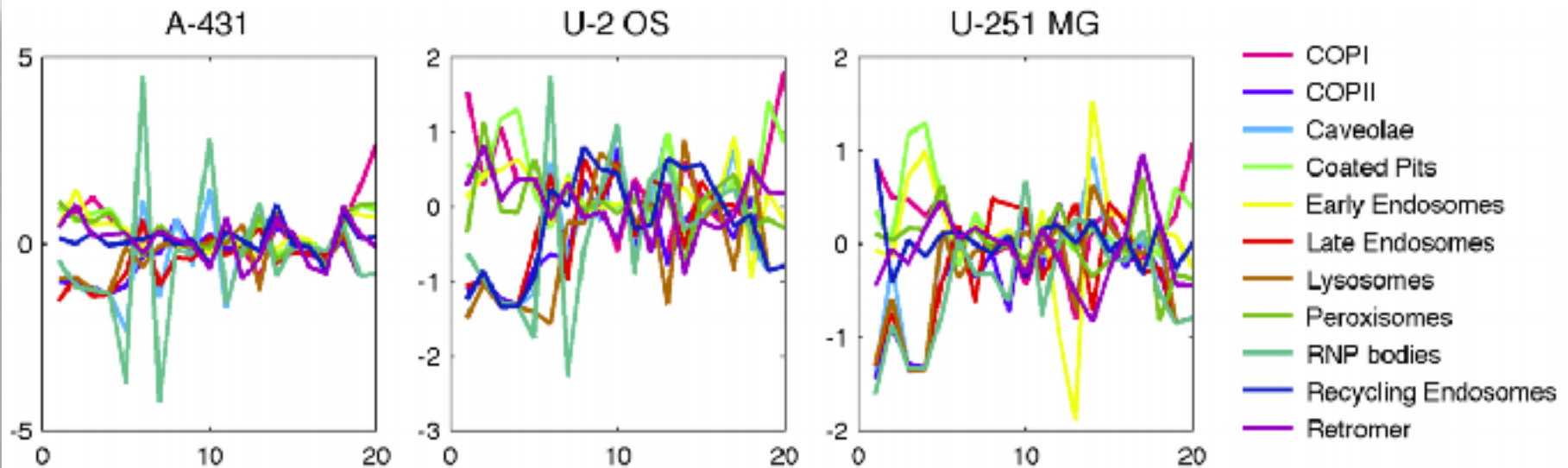
- Isolation of high spatial-frequency **Foreground (Puncta)** and **Background (Fluorescence)**
- **Compute puncta Characteristics** and **microtubule-puncta distances**
- **Probability Density Functions** for position of puncta and background intensity



# THE METHOD

## FEATURE VARIANCE

- Feature Characterization of puncta within cell regarding
  - Microtubule association / distance
  - Relationship to cell geometry
  - Density
  - Intensity
  - Appearance
- Gives rise to Feature Vector containing Major Modes of Variation among punctate patterns

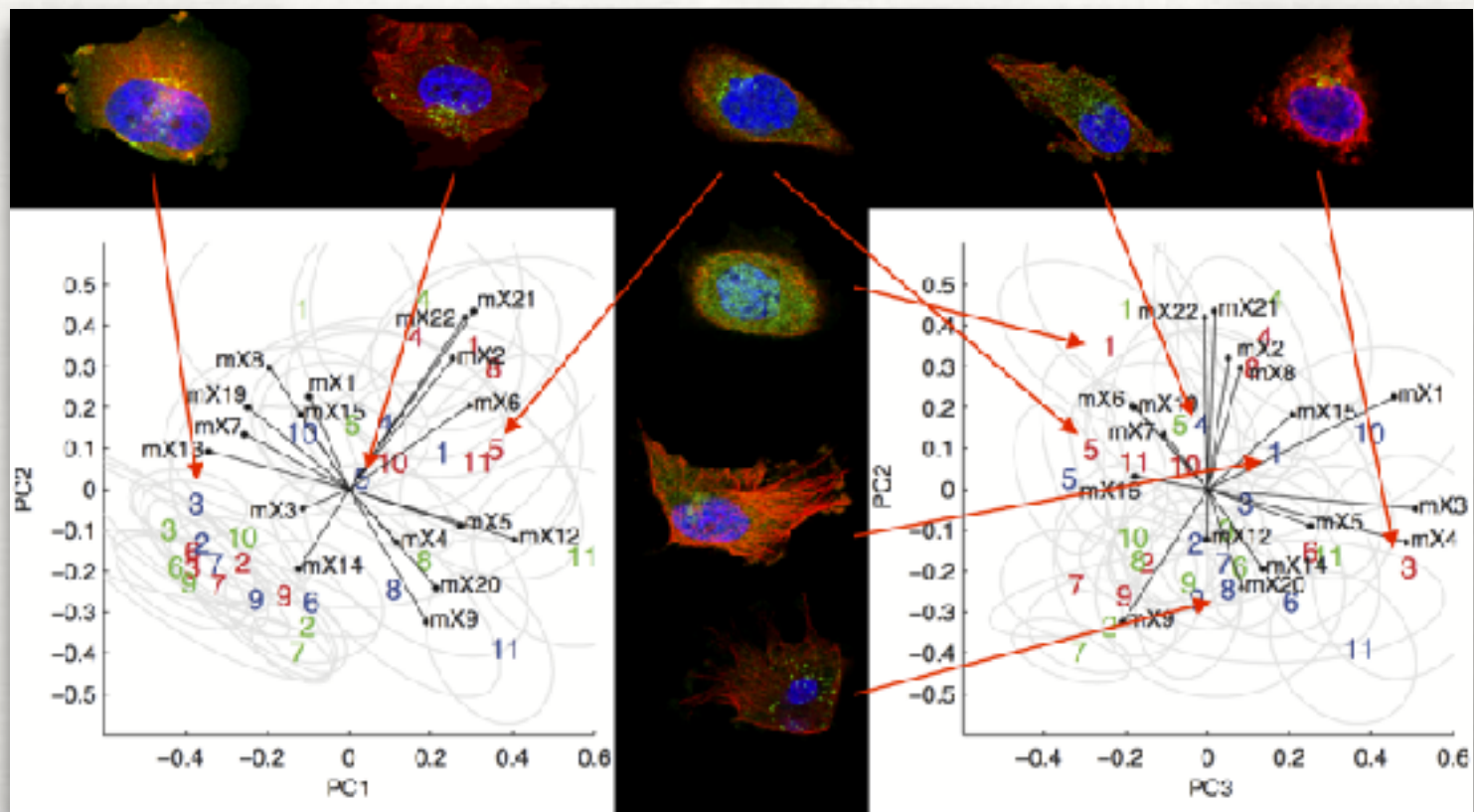




# THE METHOD

## PRINCIPAL COMPONENT ANALYSIS

- PCA shows **Variation in Features** to verify as part of reliable feature set
  - Principal components = underlying structure in the data
  - Variation in regards to principal component baseline



# THE METHOD

## CLASSIFICATION TASK

- Classification approach based on **SVM** and **Bayes Error Rate**
- **Dissimilarity Measure:** Comparing two images of cell patterns, classifying from 0 (totally inseparable) to 1 (totally separable)
- **Training Set of 11 Founder Patterns:**
  - Classification using held-out images
  - Classification only after reaching statistically significant **Threshold** (~0.72)
  - Average class accuracy = **86.9%**
  - Average class accuracy **without** microtubule distribution = **82.8%**
- Relationship to microtubules **Provides Essential Information** for Pattern

# THE RESULTS

## CLASSIFICATION

- **Test Set** of proteins other than founders:
  - Measure dissimilarity between image and each founder pattern and choose lowest
  - **“Ambiguous”** if several below multiple thresholds simultaneously
  - **125 Protein Patterns Identified**
- Found literature supporting most annotations
- No Assignment for 3 reasons:
  - Low-Quality Staining
  - Cytoplasmic proteins without discernible punctate pattern
  - Multi-pattern proteins

Gene Name	Proposed Annotation
NME6	COPII
PDE8A	Caveolae
SIAE	Early Endosomes
BRC4	Lysosomes
MMP3	RNP bodies
IARS	Retromer
ZNF155	COPII
SERPINA4	Caveolae
RAB5C	Early Endosomes
PHB	Late Endosomes
PDZK1IP1	Lysosomes
FBXD15	RNP bodies
DTX3L	Recycling Endosomes
SEPT1	Retromer
ACRC	COPI
MDGA2	COPII
LY9K	Caveolae
HCC3	Late Endosomes
CT3H	Lysosomes
NDRG4	RNP bodies

# THE METHOD

## GENERATIVE MODEL OF PUNCTATE PROTEIN DISTRIBUTIONS

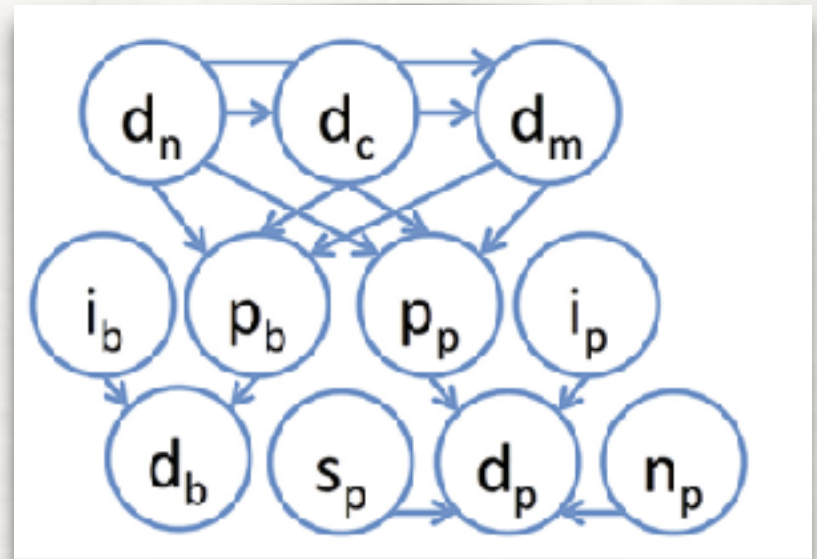
- How to best describe sub-cellular pattern?
- Current methods:
  - Descriptions using **Unstructured Text**  
—> Word insufficient for reader to mentally construct pattern
  - Show **Example Image**  
—> No information about variation
  - **Descriptive Feature** vector or matrix  
—> Only recognizes new example, does not produce example of pattern
  - None *in silico*  
—> Required for mathematical simulations of cell biochemistry and behavior
- Answer: Generative Models ?
- Capture **Underlying Properties as Statistical Distributions** to synthesize new images



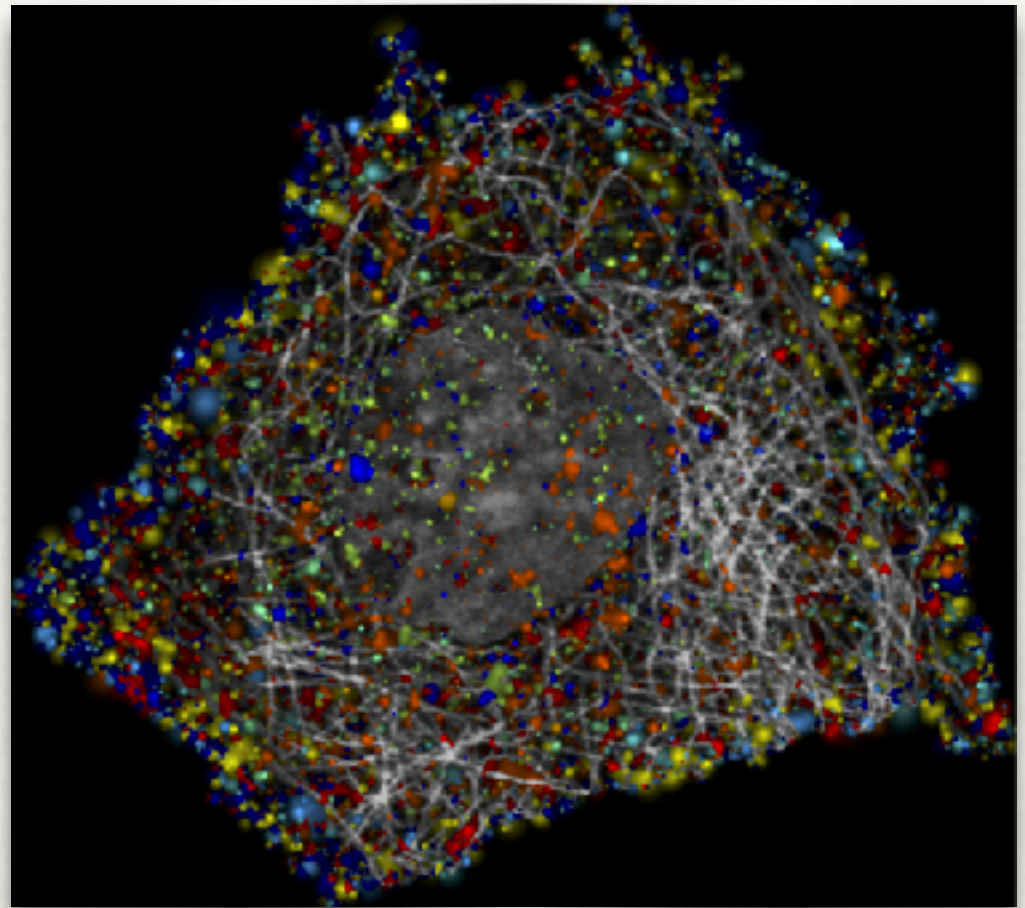
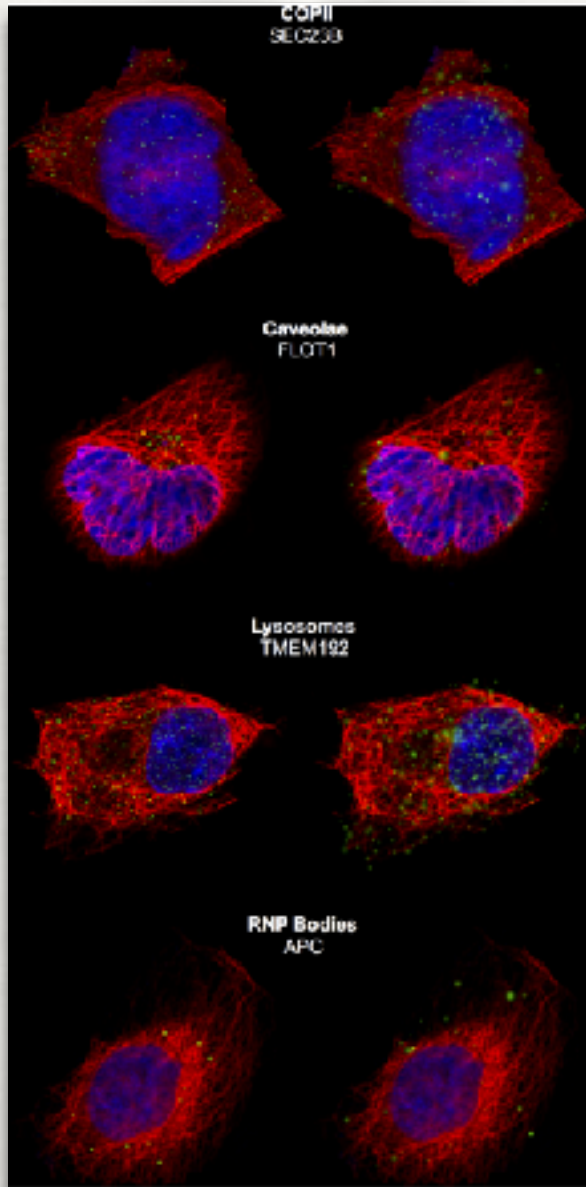
# THE METHOD

## GENERATIVE MODEL OF PUNCTATE PROTEIN DISTRIBUTIONS

- **Models of Distribution  $d$**   
(nuclear, cell shape, microtubule)
- **Models of Puncta Distribution  $p$**   
(using features capturing cell shape and microtubule dependence)
- Size shape and Intensity of vesicles modeled independently
- **Generates distributions for Foreground and Background**
- **Dependent on correct previous models**
- **Gave Rise to Fairly Accurate Image Generation**




# THE RESULTS



# DISCUSSION

- Negative
  - Paper Lacks Focus
  - Continuation of Glory & Murphy's 2007 "Automated Subcellular Location Determination and High-Throughput Microscopy"
    - Should be read in tandem with that paper
  - Readership Expectation skewed (Murphy writing papers since 1998)
  - Account for Protein Isoforms
- Positive
  - Huge Application Potential
  - Murphy Lab and [CellOrganizer.org](http://CellOrganizer.org)
  - Good Scientific Method (PCA, NOVA)

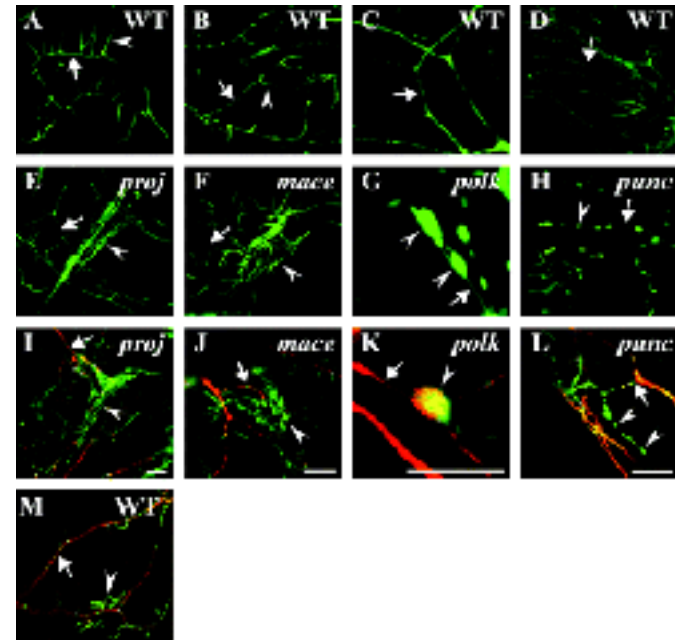
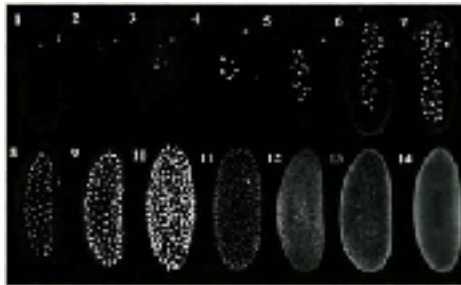
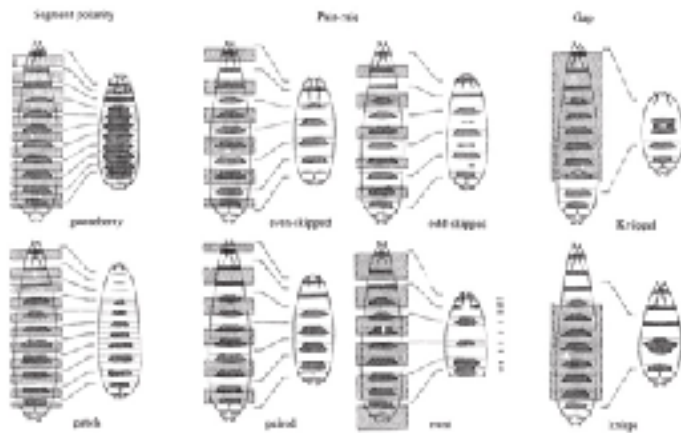
# Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning



Thouis R. Jones, Anne E. Carpenter, Michael R. Lamprecht, Jason Moffat, Serena J. Silver, Jennifer K. Grenier, Adam B. Castoreno, Ulrike S. Eggert, David E. Root, Polina Golland, and David M. Sabatini

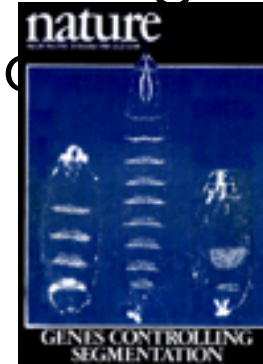
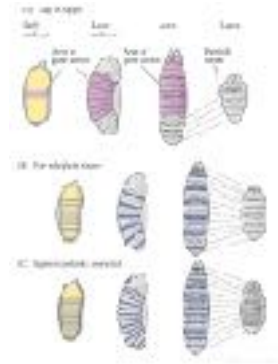
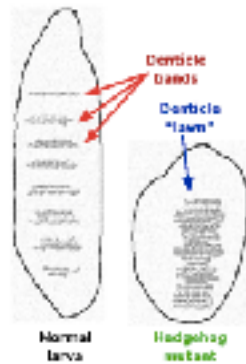


# Visual Inspection is Important for Biology!



# Why this is important

Biologists have discovered many important pathways because they found mutant organisms and decided to determine the genes that



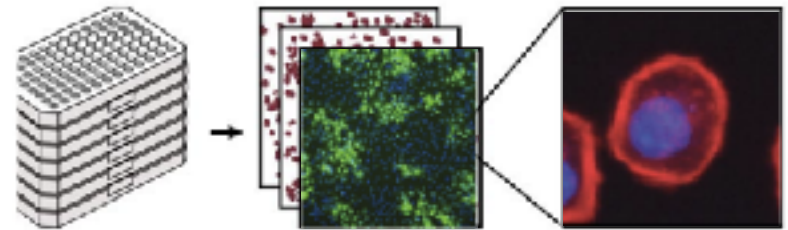
# Motivation

## Identifying mutant cells in microarrays by

“... difficult

However, analyses that cannot be achieved with the existing applications in commercial software remains challenging<sup>31</sup>. Some investigators have turned to tedious manual inspection of images for scoring: example phenotypes

identified cells in metaphase by empirically applying sequential gates based on 4 measured features of the DNA stain of each cell. This process took more than a week. With our new approach, we identified metaphase nuclei and accurately scored the entire screen within 4 h, of which only 1 h was hands-on time (Fig. S7

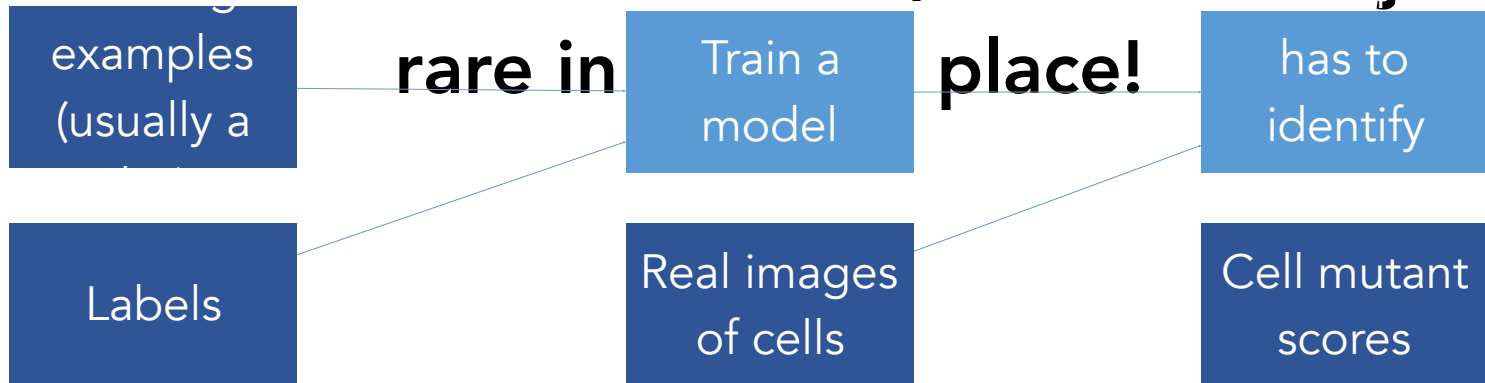


Imagine going through a huge microarray to look for individual cells!

**If we use machine learning, we often don't have training data for what mutants look like, because they are rare in the first place!**

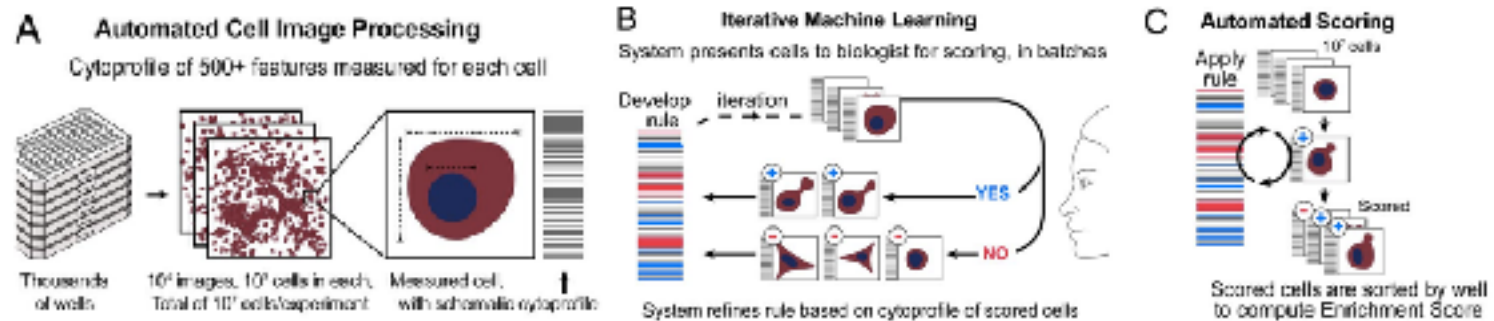
# Motivation

If we try to use machine learning to identify mutant cells, we often don't have training data for what mutants look like, because they are





# Their Solution

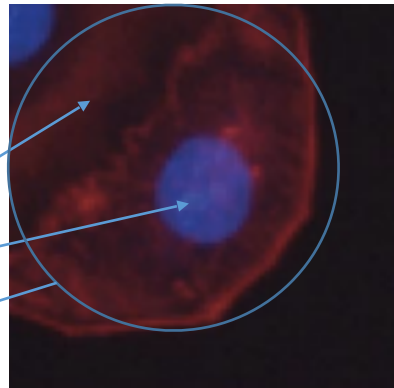
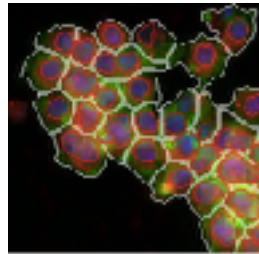


A system to automate cell image processing, perform human-in-the-loop machine learning, and automate scoring afterwards for a phenotype we are trying to identify

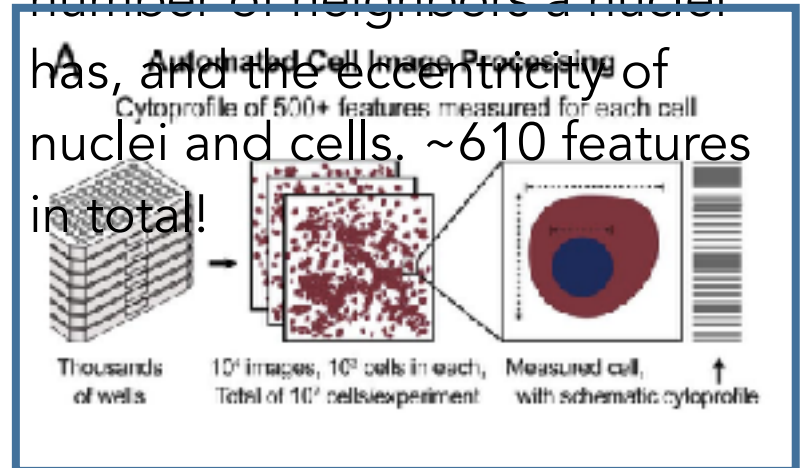
# A: Automated Cell Image Processing

**1.** Load microarray images into the pipeline. They use CellProfiler to

**2.** CellProfiler segment each cell then takes each cell in the microarray and determines the Texture, Intensity, and Shape of the nucleus, cytoplasm, and cell in general

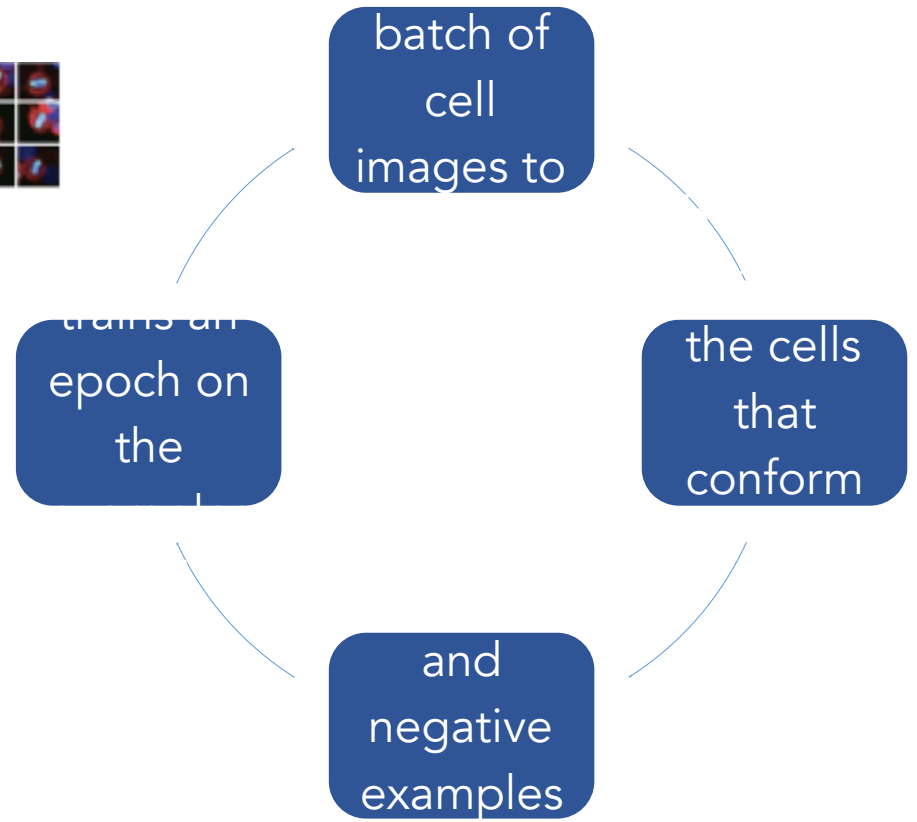
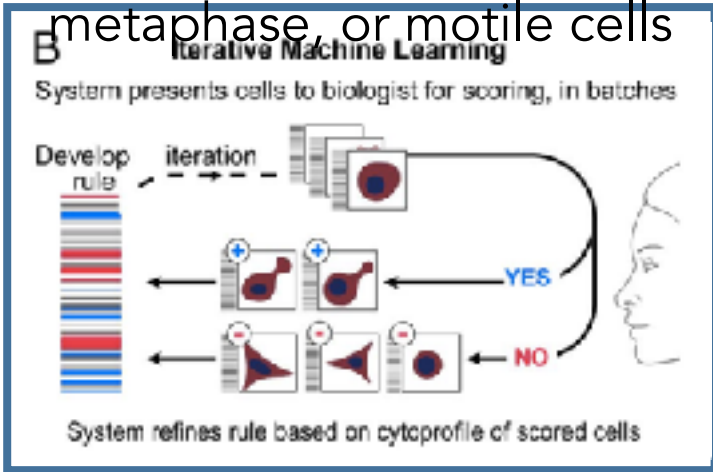
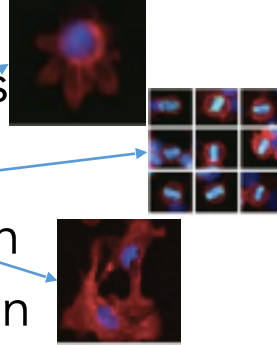


**3.** Other features are also extracted such as the number of neighbors a cell has, the number of neighbors a nuclei has, and the eccentricity of nuclei and cells. ~610 features in total!



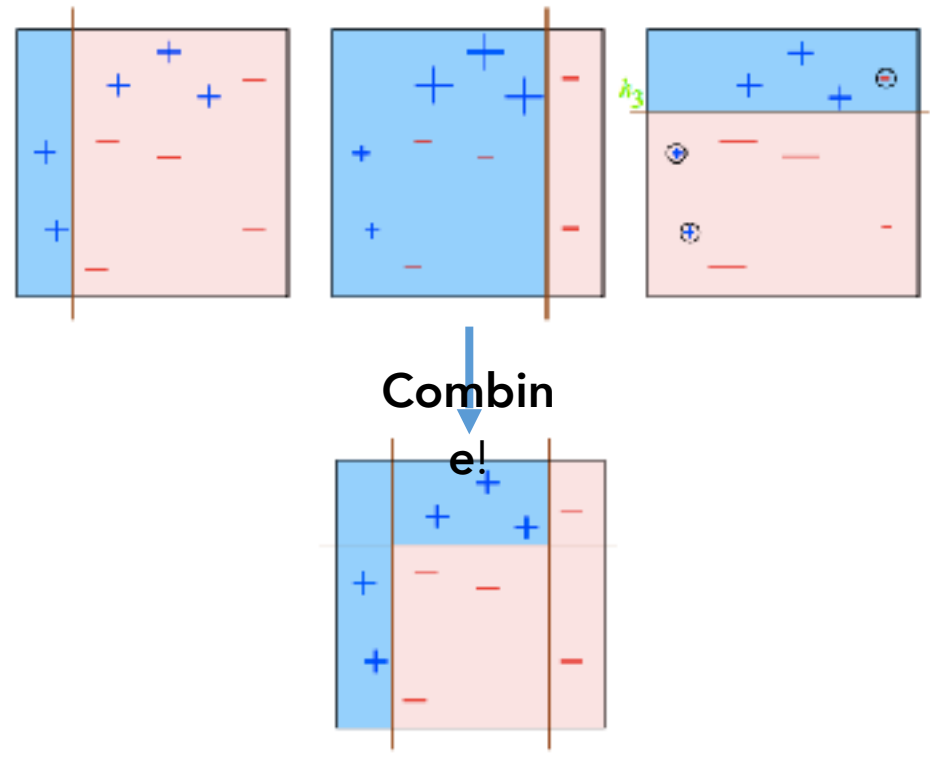
# B: Iterative Machine Learning

**First:** Biologist chooses the phenotype to identify. Examples include cells with actin blebs, cells currently in metaphase, or motile cells



# B: What is Boosting?

Old machine learning algorithm [Freund and Schapire '95] that trains many weak (dumb) classifiers to learn simple rules using coordinate descent, and combines the rules to generate more intelligent predictions. Whenever a batch is scored by the biologist, the boosting algorithm learns a new rule that splits the data and incorporates it into the model.

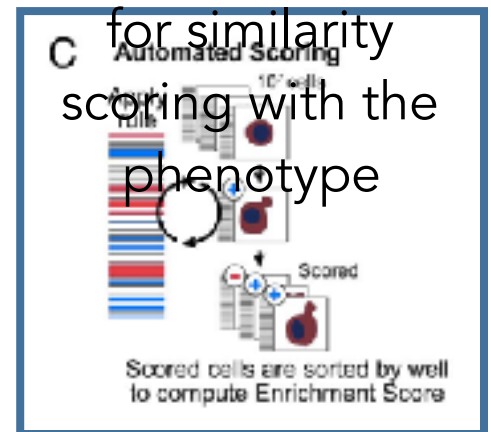


# C: Automated Scoring

Phenotype	Average Human	Human 1		Human 2		Human 3		Computer	
		Hit	Miss	Hit	Miss	Hit	Miss	Hit	Miss
Actin Dots	Hit	100	0	100	0			100	0
	Miss	0	100	0	100			0	100
Peripheral Actin	Hit	100	0	100	0			100	0
	Miss	0	100	0	100			0	100
Anaphase/Telophase	Hit	91.6	8.4	89.9	10.1			94.9	5.5
	Miss	8.4	91.6	10.1	89.9			5.5	94.9
Angular Cell Edges	Hit	99.5	0.5	98.6	1.4			99.5	0.5
	Miss	0.5	99.5	1.4	98.6			0.5	99.5
Crescents	Hit	93.3	6.7	91	9	92.6	7.4	94.2	5.8
	Miss	6.7	93.3	9	91	7.4	92.6	5.8	94.2
Prophase	Hit	96.6	3.4	94	6			97.5	2.5
	Miss	3.4	96.6	6	94			2.5	97.5
Actin Blebs	Hit	98.1	1.9	96.8	3.2			98.6	1.4
	Miss	1.9	98.1	3.2	96.8			1.4	98.6
Large Spread Cells	Hit	99.5	0.5	98.6	1.4			99.5	0.5
	Miss	0.5	99.5	1.4	98.6			0.5	99.5
Metaphase	Hit	99	1	97.3	2.7			99	1
	Miss	1	99	2.7	97.3			1	99
Motile	Hit	100	0	100	0			100	0
	Miss	0	100	0	100			0	100
Lang Projections	Hit	100	0	100	0			100	0
	Miss	0	100	0	100			0	100
Peas in a Pod	Hit	98.6	1.4	99.5	0.5			99.5	0.5
	Miss	1.4	98.6	0.5	99.5			0.5	99.5
Prometaphase	Hit	99.3	0.7	99.7	0.3			99.7	0.3
	Miss	0.7	99.3	0.3	99.7			0.3	99.7
Phospho-Histone H3 Dots	Hit	100	0	100	0			100	0
	Miss	0	100	0	100			0	100
Drosophila metaphase	Hit	96.4	3.6	94.6	5.4			97.5	2.5
	Miss	3.6	96.4	5.4	94.6			2.5	97.5

The computer does really well compared to humans!

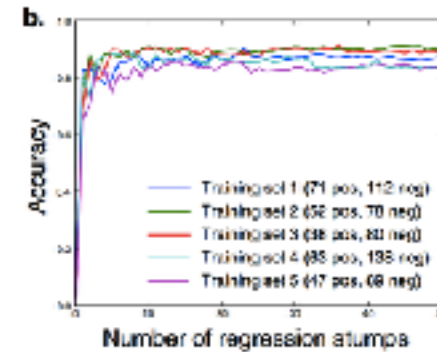
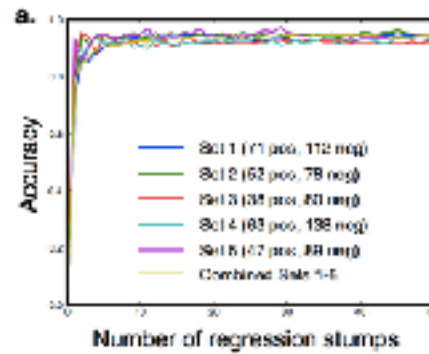
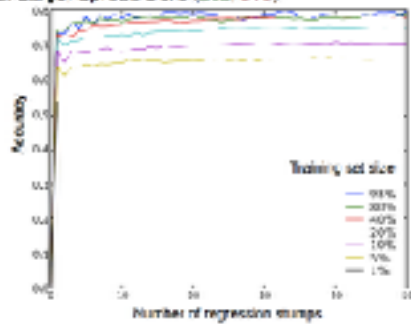
The boosting algorithm is applied on the rest of the cells



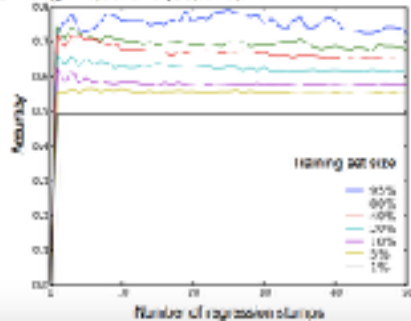


# Results: Performance vs Regression Stumps

a. Large Spread Code (202, 376)



b. Long Projections (58, 345)



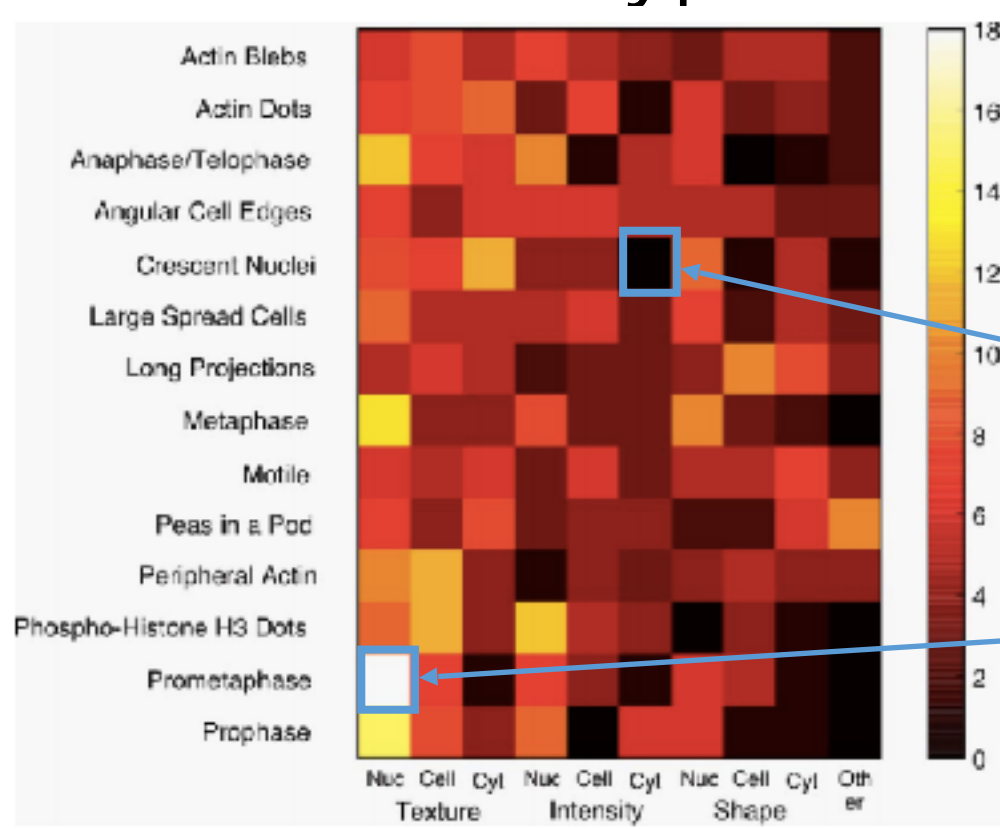
More regression stumps = higher accuracy. Curves tend to look like this when things go well:

Accuracy  
y



Regression  
stumps

# Results: Phenotype vs Feature Power



The lighter the square, the more important that feature is to a phenotype

cytoplasm intensity is not very important to identifying crescent nuclei

Nuclear texture is very important to identifying prometaphase



# Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data

Fernando Amat, William Lemon, Daniel P Mossing, Katie McDole, Yinan Wan, Kristin Branson, Eugene W Myers, & Philipp J Keller

---

PRESENTATION BY PAVITRA RENGARAJAN

CS 371 LECTURE 3/6

# Motivation

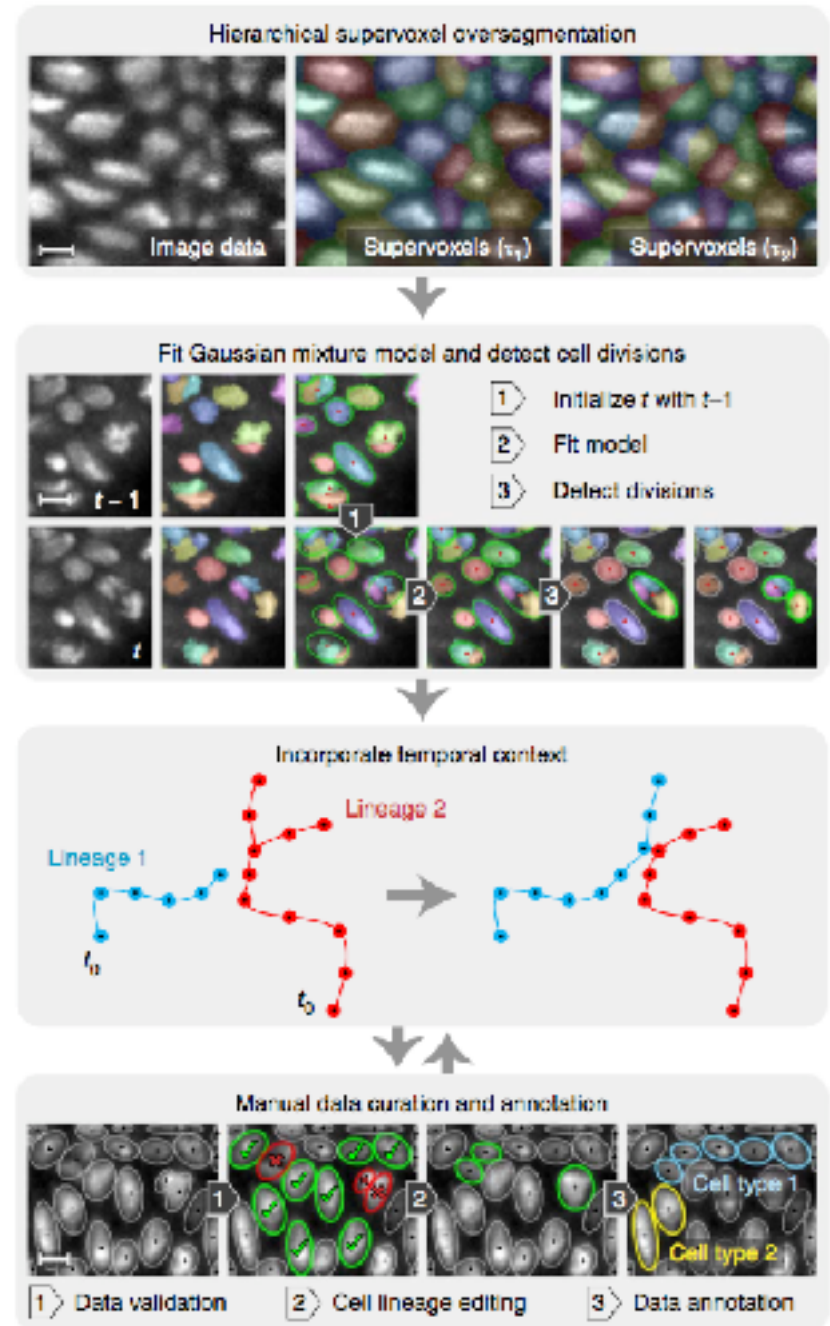
- **Cell lineage:** developmental history of a cell as traced back to the cell from which it arises
- **Cell lineage reconstruction:** accurate reconstruction of the positions, movements, and divisions of cells
  - Important goal for developmental biology
- **Computational methods** for cell lineage reconstruction involve state-of-the-art live imaging technologies that record development at cellular level for several days
  - Yields terabytes of data
- Automated cell **segmentation** (identifying cells in an image) & **tracking** (following cell movement over time) is challenging, capturing divisions is even more challenging



# Pipeline Overview

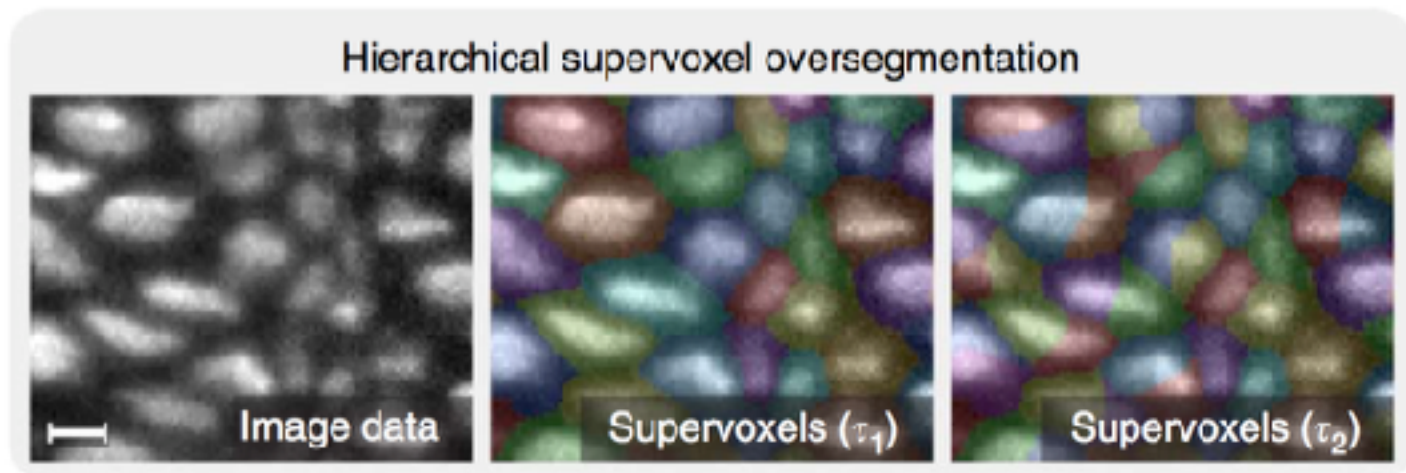
New approach to automated cell lineage reconstruction:

- 1) Segment image by identifying individual cells
- 2) Detect cell divisions using a probabilistic model
- 3) Flag areas where model might have failed and use heuristic rules or manual inspection



# Automated Cell Segmentation

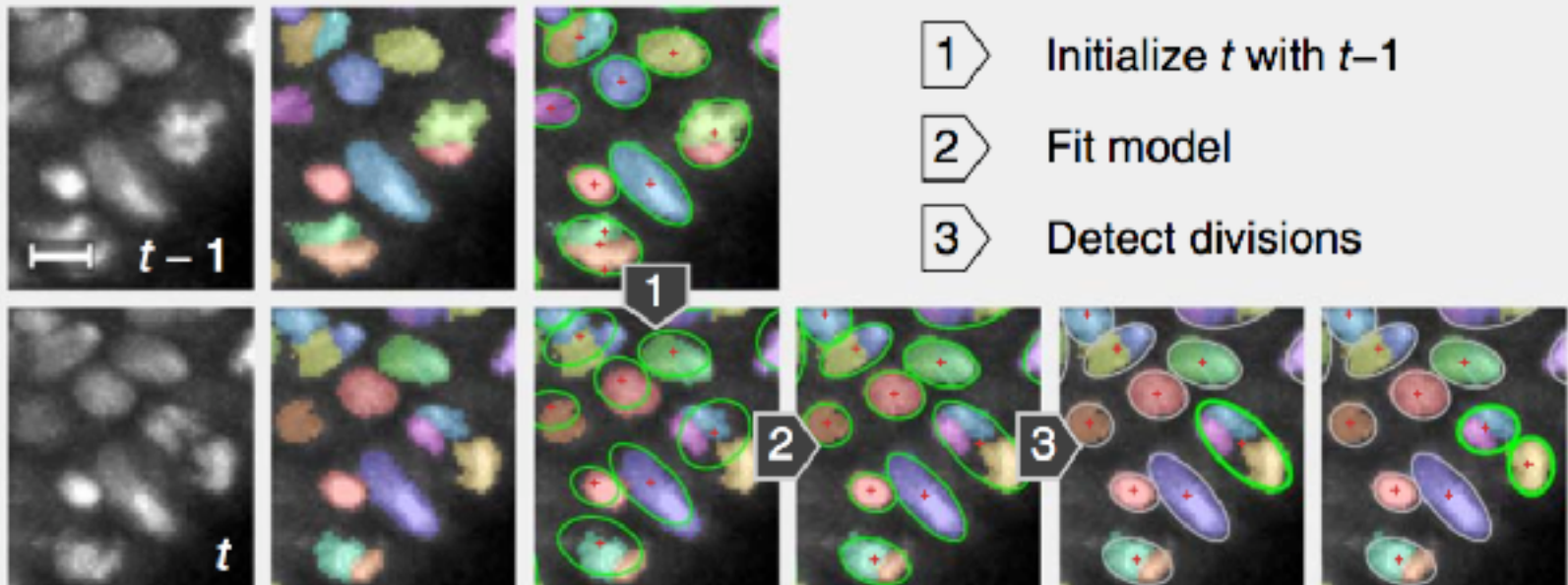
- Considered all possible partitions of image into “supervoxels”
- **Voxel:** small unit that defines a point in 3D space
- **Supervoxel:** connected set of voxels belonging to a nucleus; each nucleus can be represented by multiple supervoxels
- Requires only 2 parameters:  $\tau$ , which affects merging of image regions, and global background intensity threshold



# Connecting Supervoxels in Space/Time

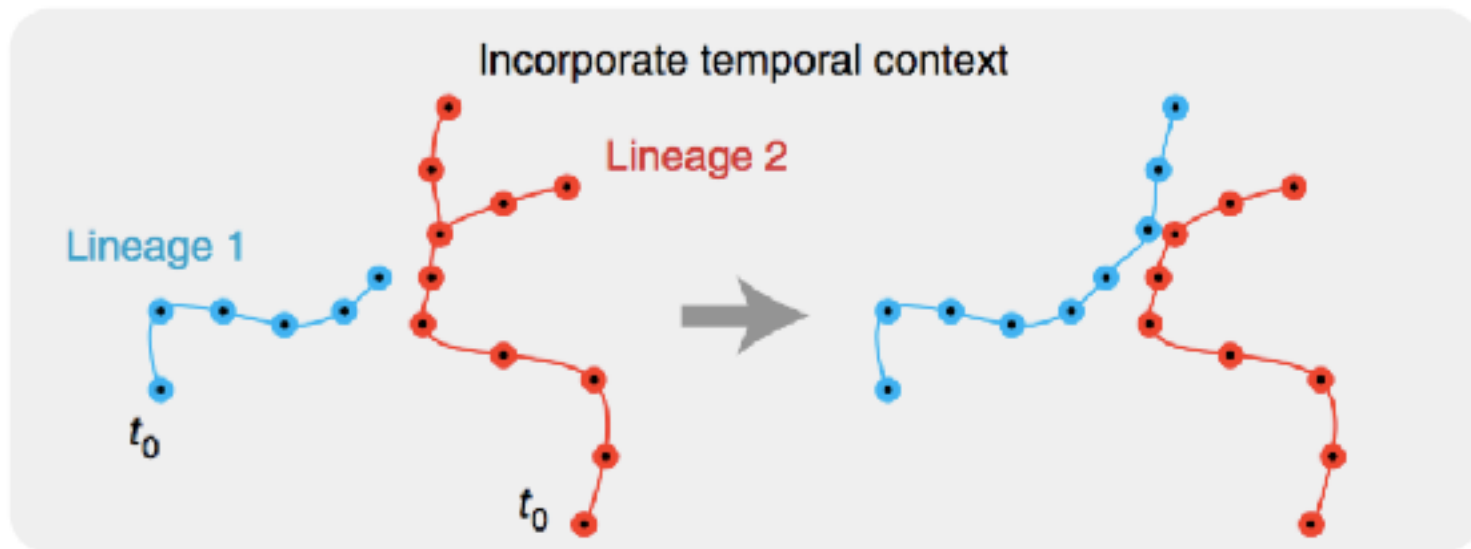
- Modeled cell location by nucleus-specific fluorescent labels
- Detected cell divisions by using probabilistic model called Gaussian mixture model

Fit Gaussian mixture model and detect cell divisions



# Potential Failure Flagging

- Apply heuristic rules to improve accuracy
- Algorithm determines local spatiotemporal windows in which the model might have been erroneous

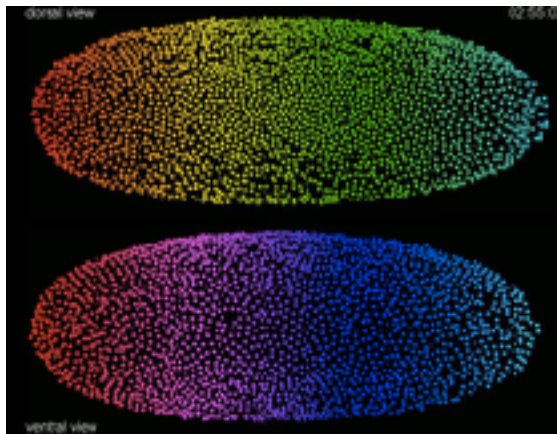




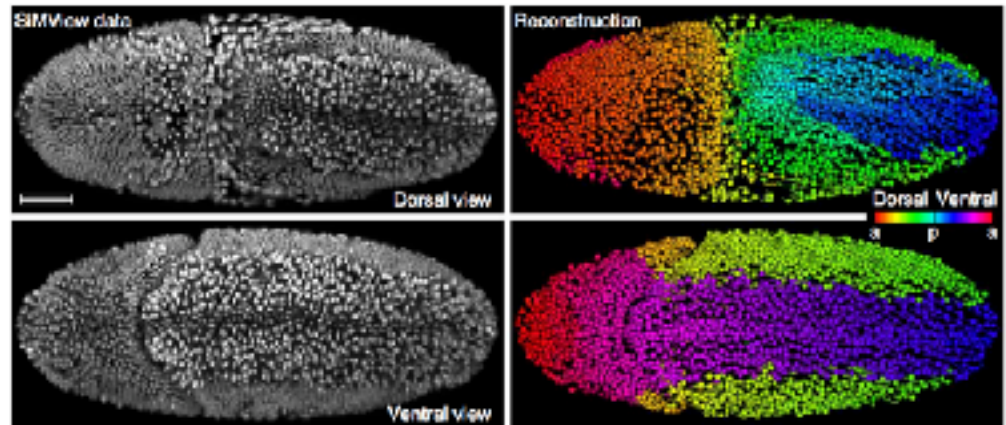
# Results

## Analysis of lightsheet microscopy on *Drosophila* embryonic development:

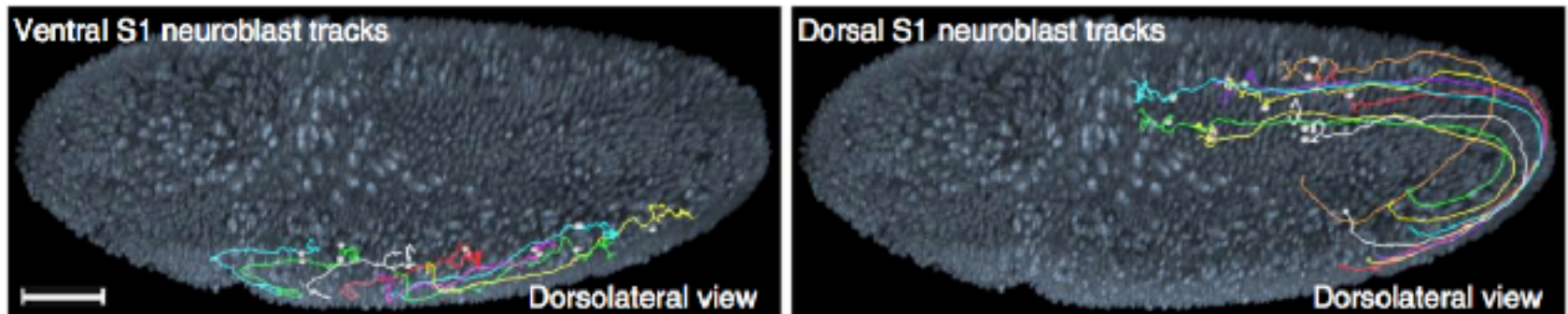
Initial:



After 24 hours:



## Tracks of eight such neuroblasts during germ band extension:





# Automated Segmentation & Tracking

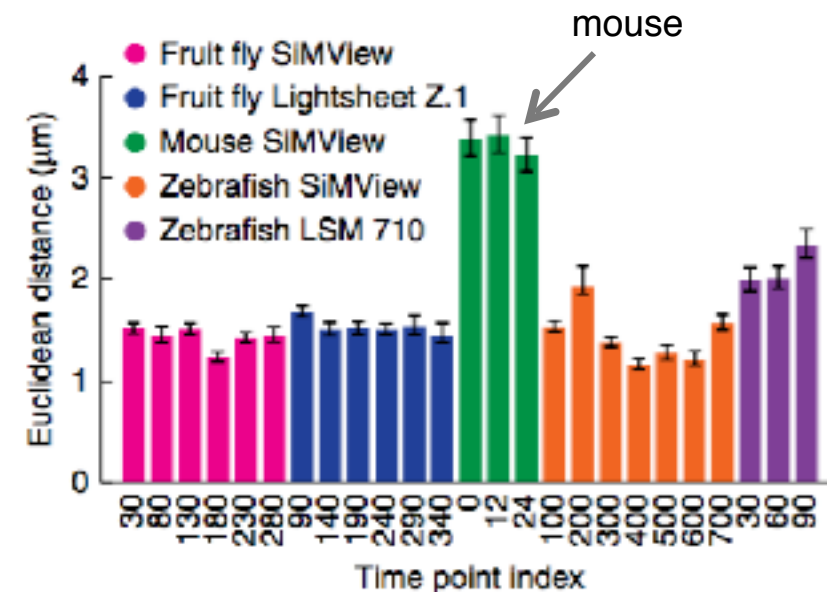


[http://www.nature.com/nmeth/journal/v11/n9/fig\\_tab/nmeth.3036\\_SV2.html](http://www.nature.com/nmeth/journal/v11/n9/fig_tab/nmeth.3036_SV2.html)

# Aggregate Results

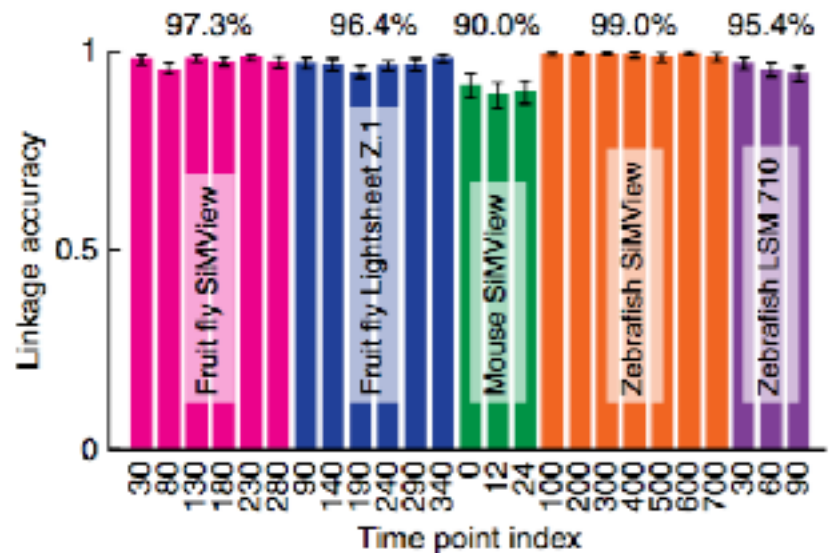
**Euclidean distance** between manually and automatically marked nuclei centroids

- Average Euclidean distance below nuclear radius



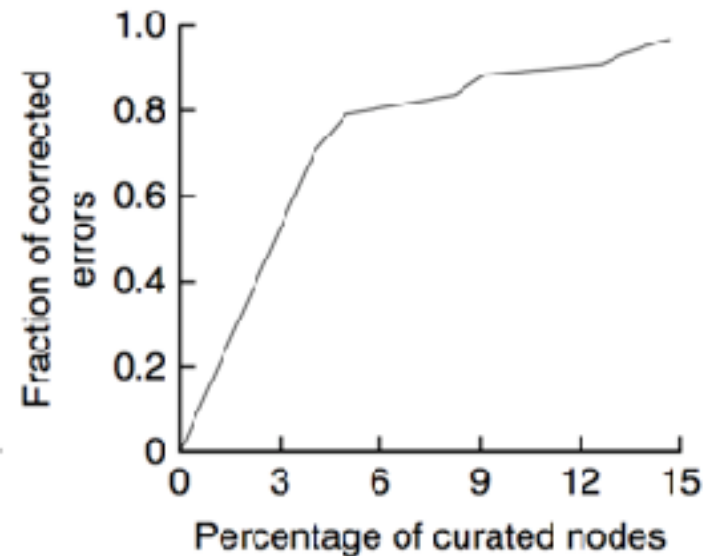
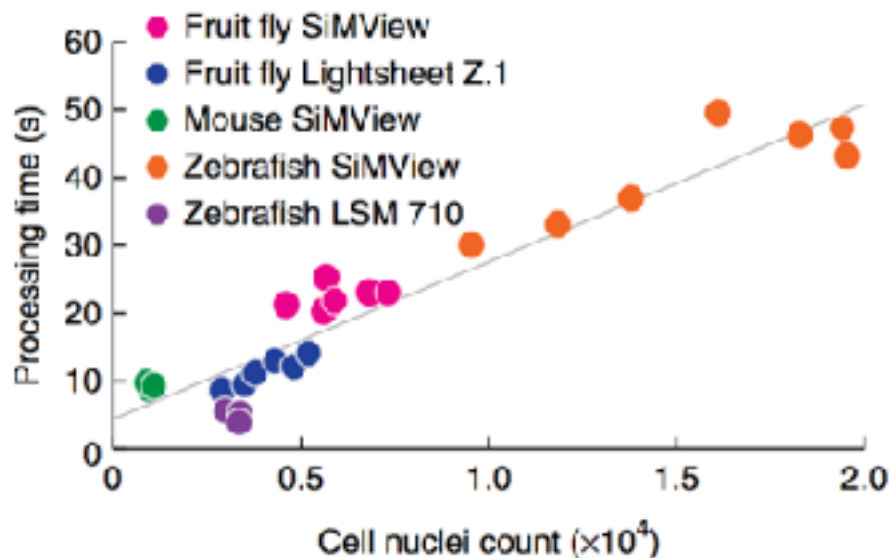
**Linkage accuracy:** fraction of correct linkage assignments in consecutive time points

- Average linkage accuracy between 90% and 99%



# Performance

- Linear scaling of computation time with the number of cells tracked when parallelizing on multicore CPU & GPU platforms
- Manual inspection of only 15% of all data points was required to correct 97% of the errors



# Recap

- Strengths
  - **Generality** – considered 3 different model types with 3 different types of fluorescence microscopes
  - **Scalability** – analyzed terabyte-sized data with up to 20,000 cells per time point at 26,000 cells per minute on single computer workstation
  - **Ease of use** – adjusted only two parameters
- Weaknesses
  - Flagged all cell divisions & cell deaths for manual inspection
  - Naïve assumption to consider context of only 1 time step
  - Could have used clearer explanation of performance gain
- Could pave the way for “**smart microscopes**”

# References

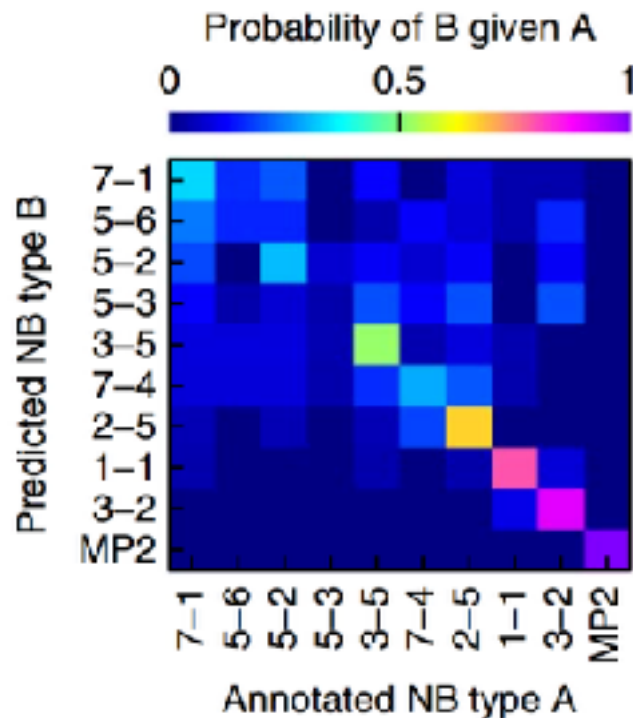
Presentation figures/content from the following paper:

**Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data.** Amat, Fernando; Lemon, William; Mossing, Daniel P; McDole, Katie; Wan, Yinan; Branson, Kristin; Myers, Eugene W; Keller, Philipp J. *Nature Methods*. Vol. 11, No. 9, September 2014.



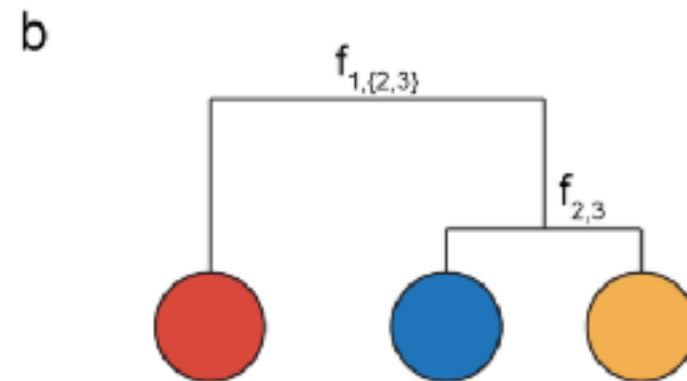
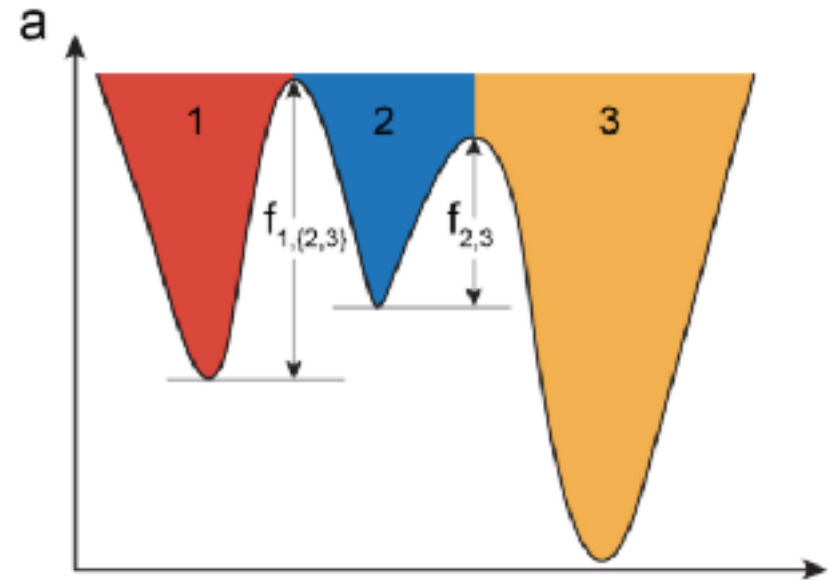
# Additional Work: Predicting Neuroblast Cell Types

- Used machine learning to predict neuroblast cell types using just information about timing and orientation of cell divisions
- Achieved 6-fold to 10-fold higher probability than probability of assigning the correct cell identity at random

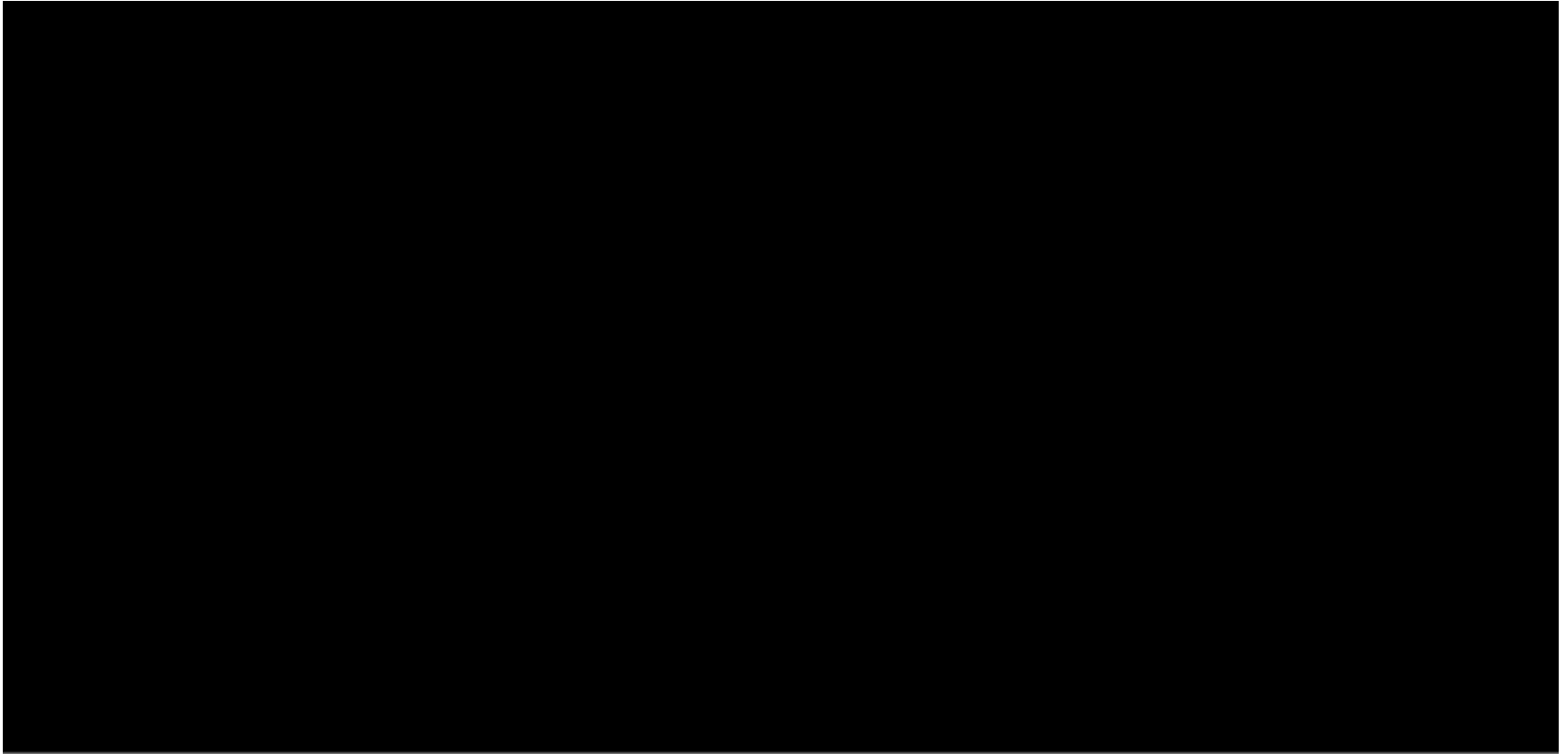


# Additional Slide: Supervoxel Partitioning Methodology

- **Algorithm:** watershed techniques and persistence-based clustering
- **Intuition:** group voxels into coherent regions belonging to the same nucleus
- Use a parameter ( $\tau$ ) to determine a hierarchical order between the basins



# Additional Slide: Drosophila Cell Lineage Reconstruction



[http://www.nature.com/nmeth/journal/v11/n9/fig\\_tab/nmeth.3036\\_SV28.html](http://www.nature.com/nmeth/journal/v11/n9/fig_tab/nmeth.3036_SV28.html)