OMERO and PSLID

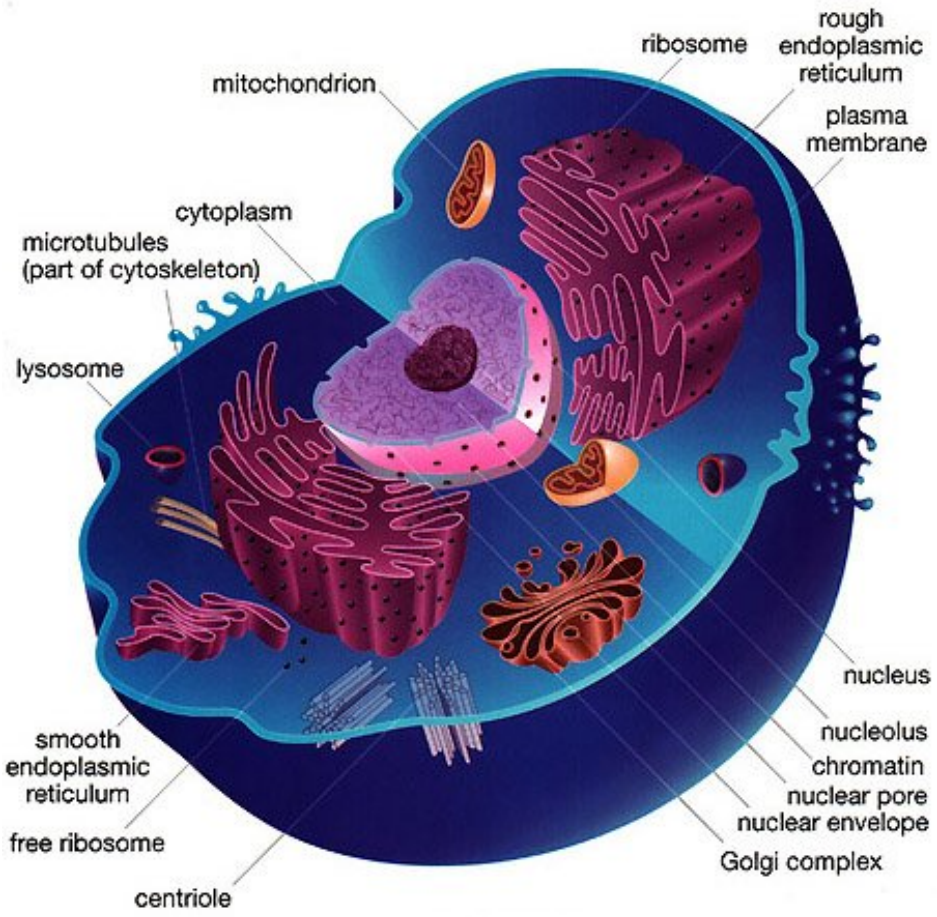
Robert F. Murphy
Lane Center for Computational Biology
Carnegie Mellon University

PSLID mission and overview

Determination of Protein Location from Microscopy Images

- Location Proteomics: Providing information about subcellular protein location
  - → (dynamic) proteome map to understand normal and abnormal cell behaviors
  - → clues for interaction/ regulation of proteins
  - → clues for protein function

- Data driven: Determination of location for all proteins from microscopy images



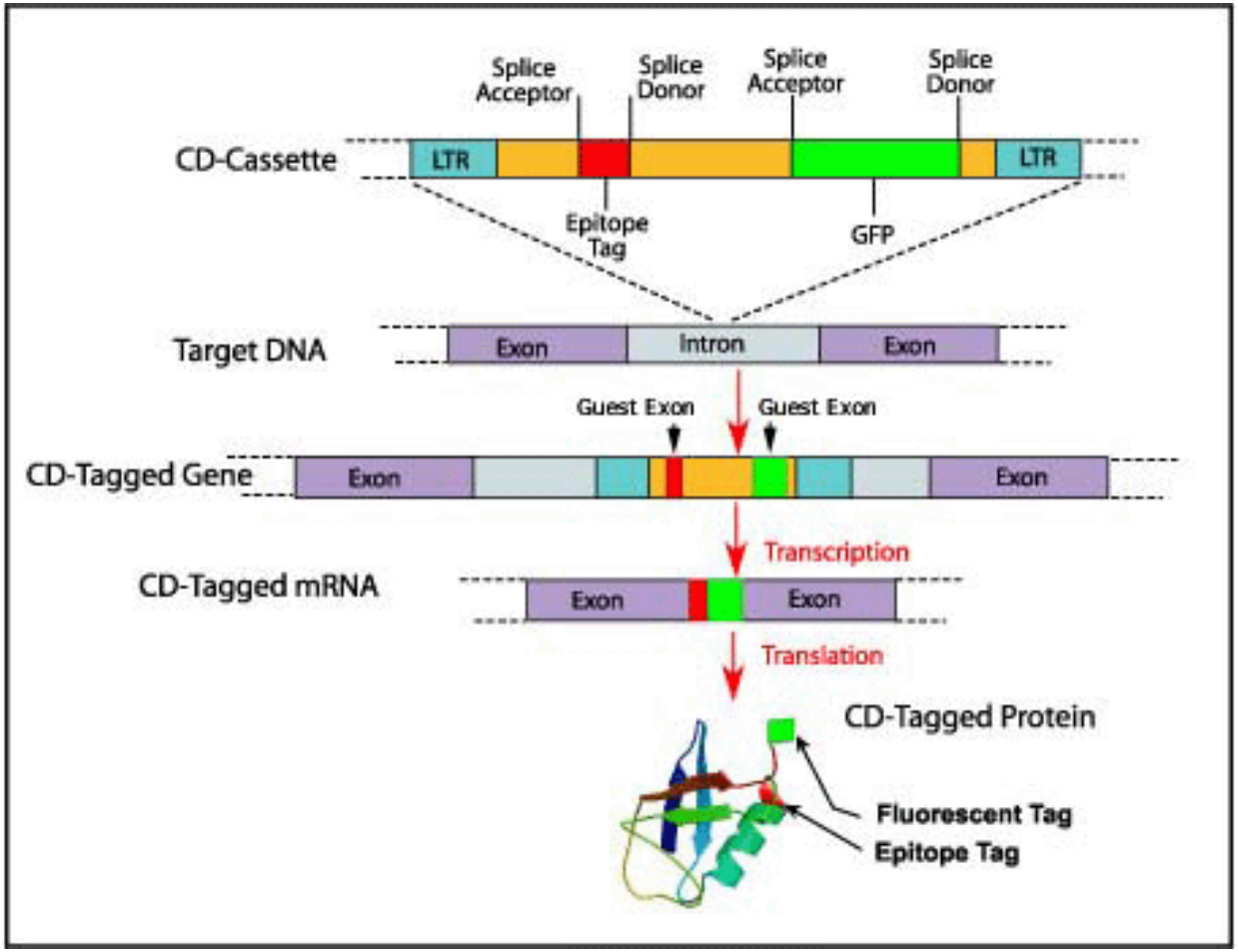
CD-tagging collection

- **Proteome-scale, live cell imaging of tagged proteins**
- Using **CD-tagging** (developed by Jonathan Jarvik and Peter Berget): Infect population of cells with a retrovirus carrying DNA sequence that will “tag” in a random gene *in its genomic context*

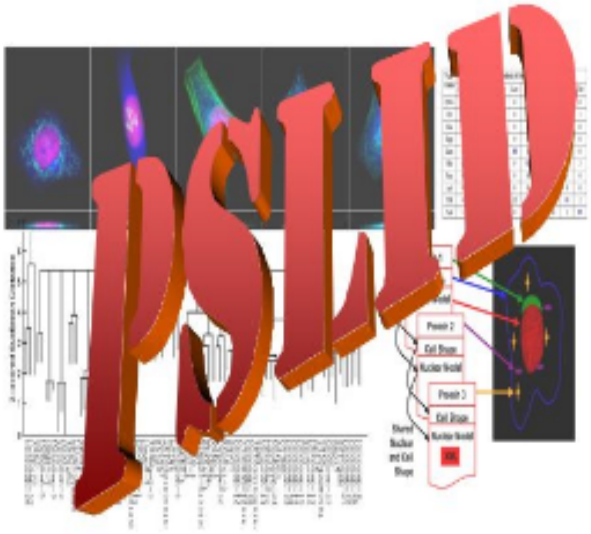
• Isolate separate clones, each of which produces express one tagged protein

• Use RT-PCR to identify tagged gene in each clone

• Collect many live cell images for each clone using fluorescence microscopy



http://pslid.org



Protein Subcellular Location Image Database

PSLID - Protein Subcellular Location Image Database

Go to the PSLID database containing CD-tagged NIH 3T3 cell clones **New!**

Go to the PSLID Public Database Release 4 (released August 1, 2008)

Download datasets contained within PSLID

Download SLIC feature calculation and pattern analysis software used in PSLID and full PSLID software

---

Supported by the National Institute of General Medical Sciences, Grant GM075205

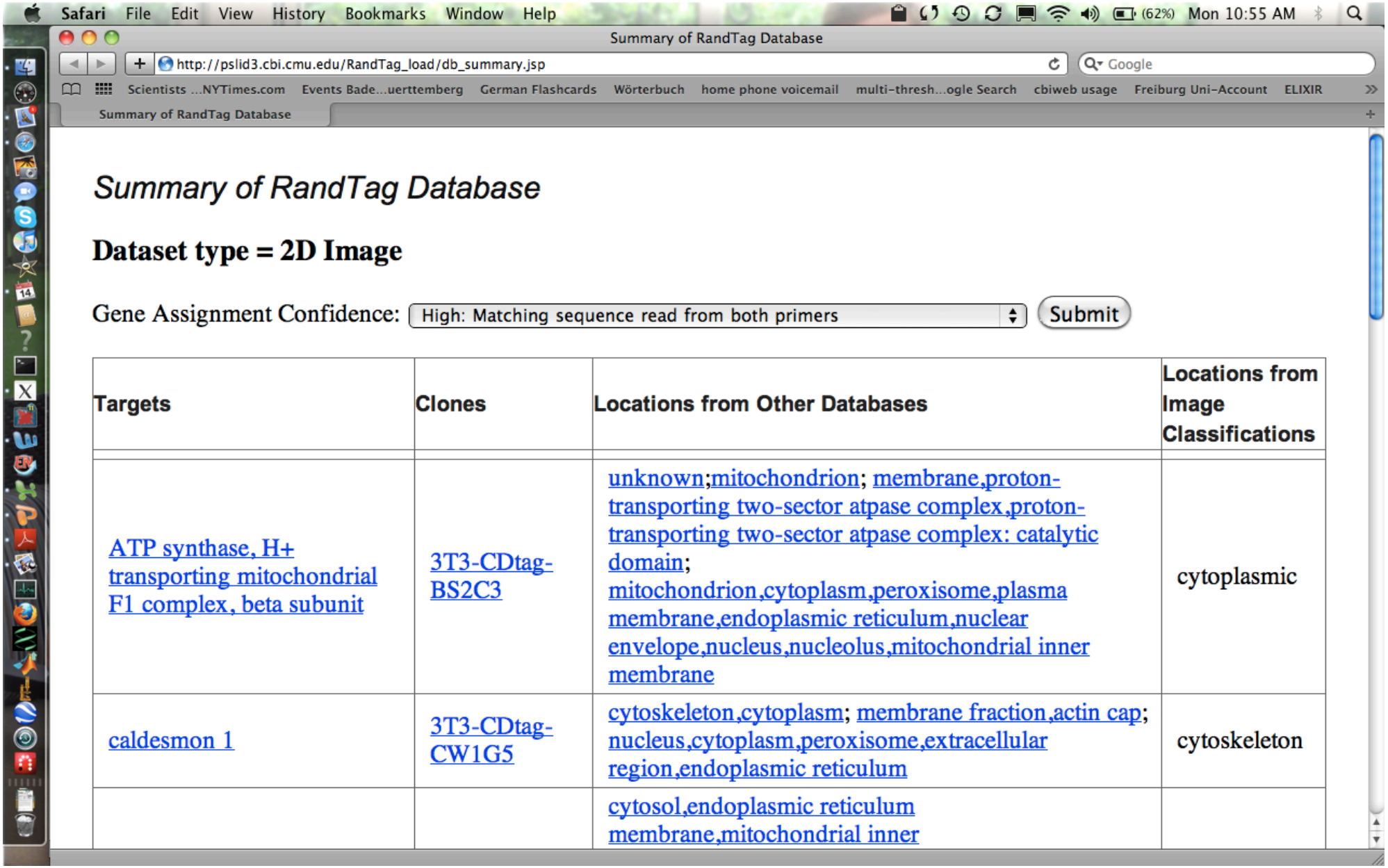
PSLID RandTag collection Release 2

- 4441 GFP-tagged 3T3 cell clones
- 15-25 images were acquired by automated microscopy (IC100) for each clone for a total of ~140,000 images. Of those, ~50,000 were classified as unusable (out of focus, debris containing, too few cells) so only ~90,000 images appear in the final sets.
- ~300 distinct proteins identified; tagged protein in remaining clones being identified by sequencing
- 667 images from confocal microscopy images for interesting clones
- Release 2 to be release on July 11, 2010 (ISMB meeting)

PSLID principles

- Fundamental unit is a set
- Elements of sets strongly typed
  - 2D, 3D, …
  - image, cell, object…
- Analysis results stored in DB, linked to element
- Meta analysis
- Search by image content

PSLID RandTag collection



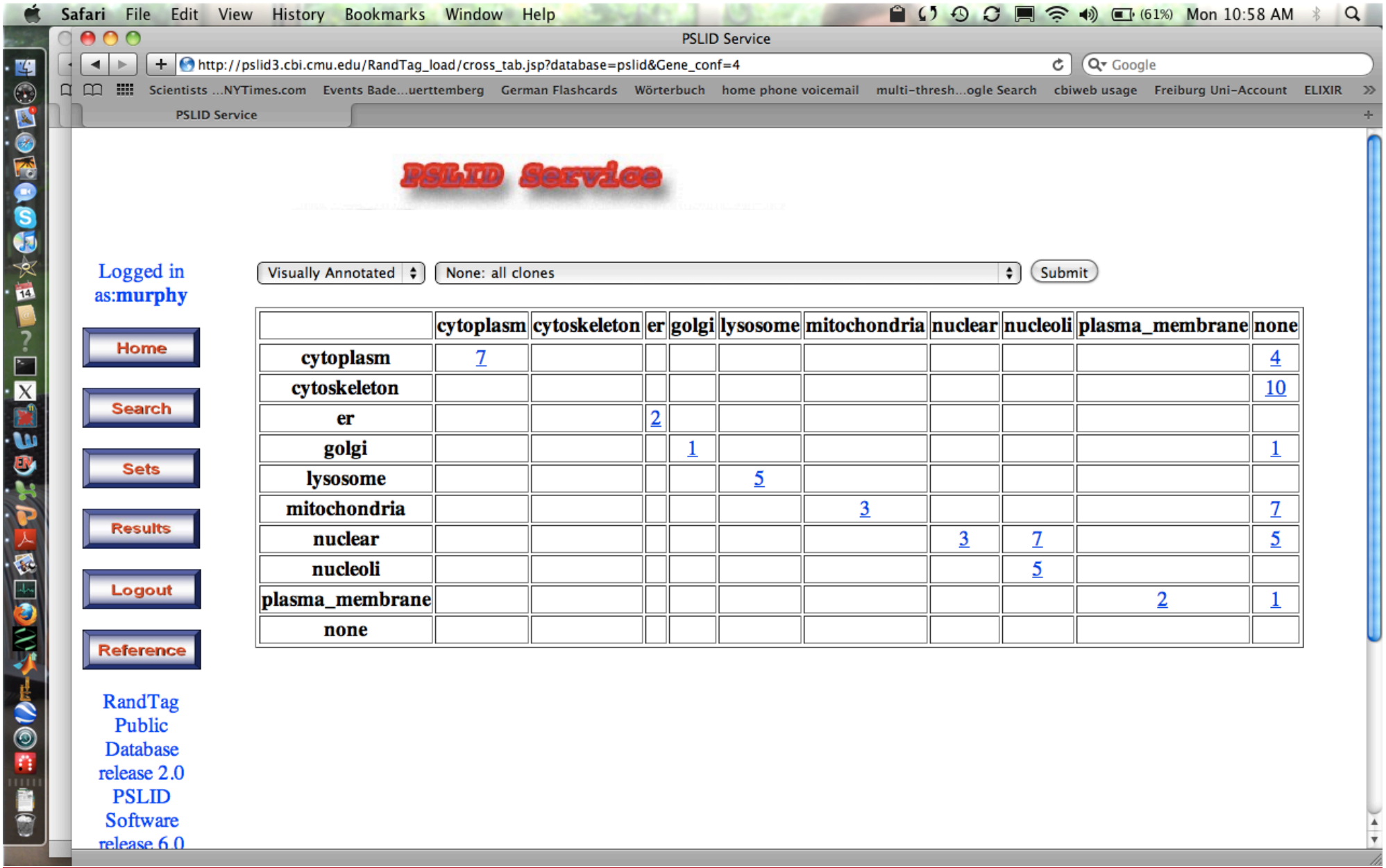
Summary of RandTag Database

Dataset type = 2D Image

Gene Assignment Confidence: High: Matching sequence read from both primers Submit

Targets	Clones	Locations from Other Databases	Locations from Image Classifications
ATP synthase, H+ transporting mitochondrial F1 complex, beta subunit	3T3-CDtag-BS2C3	unknown;mitochondrion; membrane,proton-transporting two-sector atpase complex,proton-transporting two-sector atpase complex: catalytic domain; mitochondrion,cytoplasm,peroxisome,plasma membrane,endoplasmic reticulum,nuclear envelope,nucleus,nucleolus,mitochondrial inner membrane	cytoplasmic
caldesmon 1	3T3-CDtag-CW1G5	cytoskeleton,cytoplasm; membrane fraction,actin cap; nucleus,cytoplasm,peroxisome,extracellular region,endoplasmic reticulum	cytoskeleton
		cytosol,endoplasmic reticulum membrane,mitochondrial inner	

PSLID Randtag collection



	cytoplasm	cytoskeleton	er	golgi	lysosome	mitochondria	nuclear	nucleoli	plasma_membrane	none
cytoplasm	7									4
cytoskeleton										10
er			2							
golgi				1						1
lysosome					5					
mitochondria						3				7
nuclear							3	7		5
nucleoli								5		
plasma_membrane									2	1
none										

Many ways to access and display the data

- Look at overview of the database content

→ **Database summaries (by target, by subcellular location)**

- Automate the query, to include the link in a website for example

→ **use the URL query**

By Target:
http://pslid3.cbi.cmu.edu/RandTag_develop/search.jsp?target=40S 20ribosomal 20protein 20S28

By experiment :
http://pslid3.cbi.cmu.edu/RandTag_develop/search.jsp?cell_name=3T3-CDtag-BM1B10

Returning the answer as HTML or XML format

- browsing the collection by using the web pages

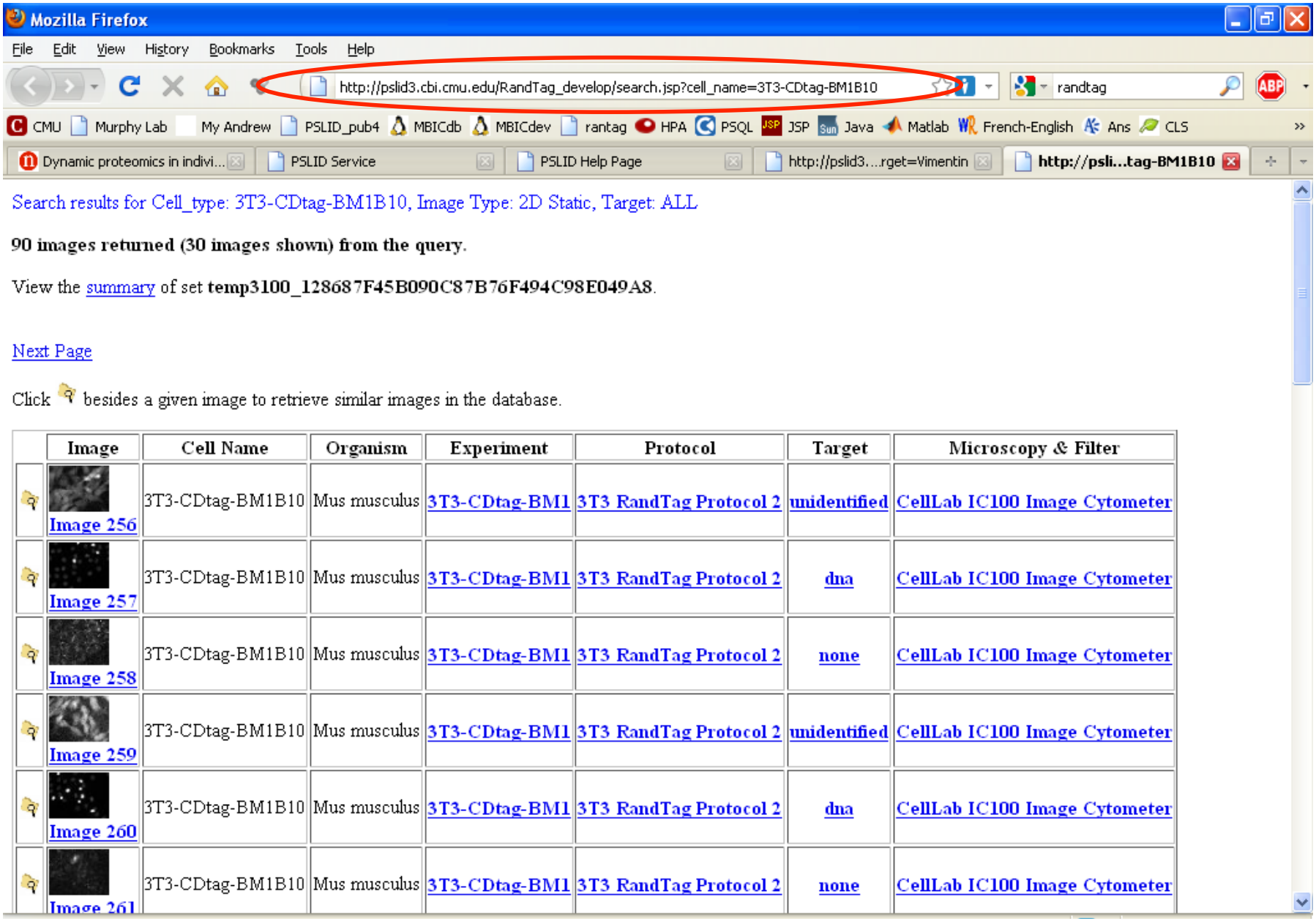
→ View **plates**

→ View Summary of the **well**

→ View all the **images** of the well

→ View **Set**

URL search: by cell type



Search results for Cell_type: 3T3-CDtag-BM1B10, Image Type: 2D Static, Target: ALL

90 images returned (30 images shown) from the query.

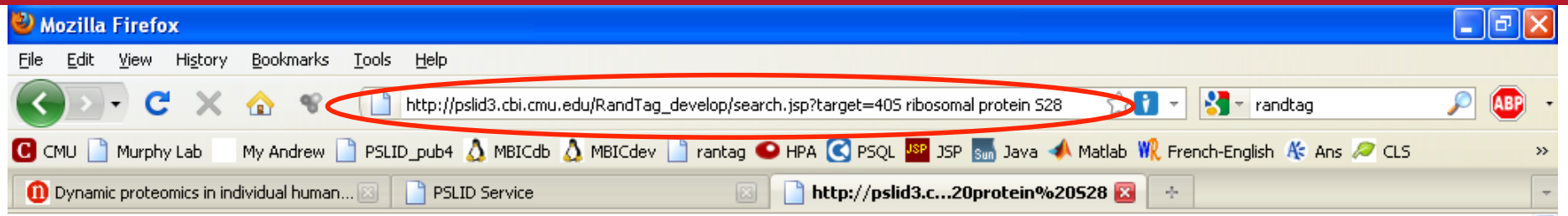
View the summary of set **temp3100_128687F45B090C87B76F494C98E049A8**.

Next Page

Click besides a given image to retrieve similar images in the database.

Image	Cell Name	Organism	Experiment	Protocol	Target	Microscopy & Filter
Image 256	3T3-CDtag-BM1B10	Mus musculus	3T3-CDtag-BM1	3T3 RandTag Protocol 2	unidentified	CellLab IC100 Image Cytometer
Image 257	3T3-CDtag-BM1B10	Mus musculus	3T3-CDtag-BM1	3T3 RandTag Protocol 2	dna	CellLab IC100 Image Cytometer
Image 258	3T3-CDtag-BM1B10	Mus musculus	3T3-CDtag-BM1	3T3 RandTag Protocol 2	none	CellLab IC100 Image Cytometer
Image 259	3T3-CDtag-BM1B10	Mus musculus	3T3-CDtag-BM1	3T3 RandTag Protocol 2	unidentified	CellLab IC100 Image Cytometer
Image 260	3T3-CDtag-BM1B10	Mus musculus	3T3-CDtag-BM1	3T3 RandTag Protocol 2	dna	CellLab IC100 Image Cytometer
Image 261	3T3-CDtag-BM1B10	Mus musculus	3T3-CDtag-BM1	3T3 RandTag Protocol 2	none	CellLab IC100 Image Cytometer

URL search: by target



Search results for Image Type: 2D Static, Target: 40S ribosomal protein S28

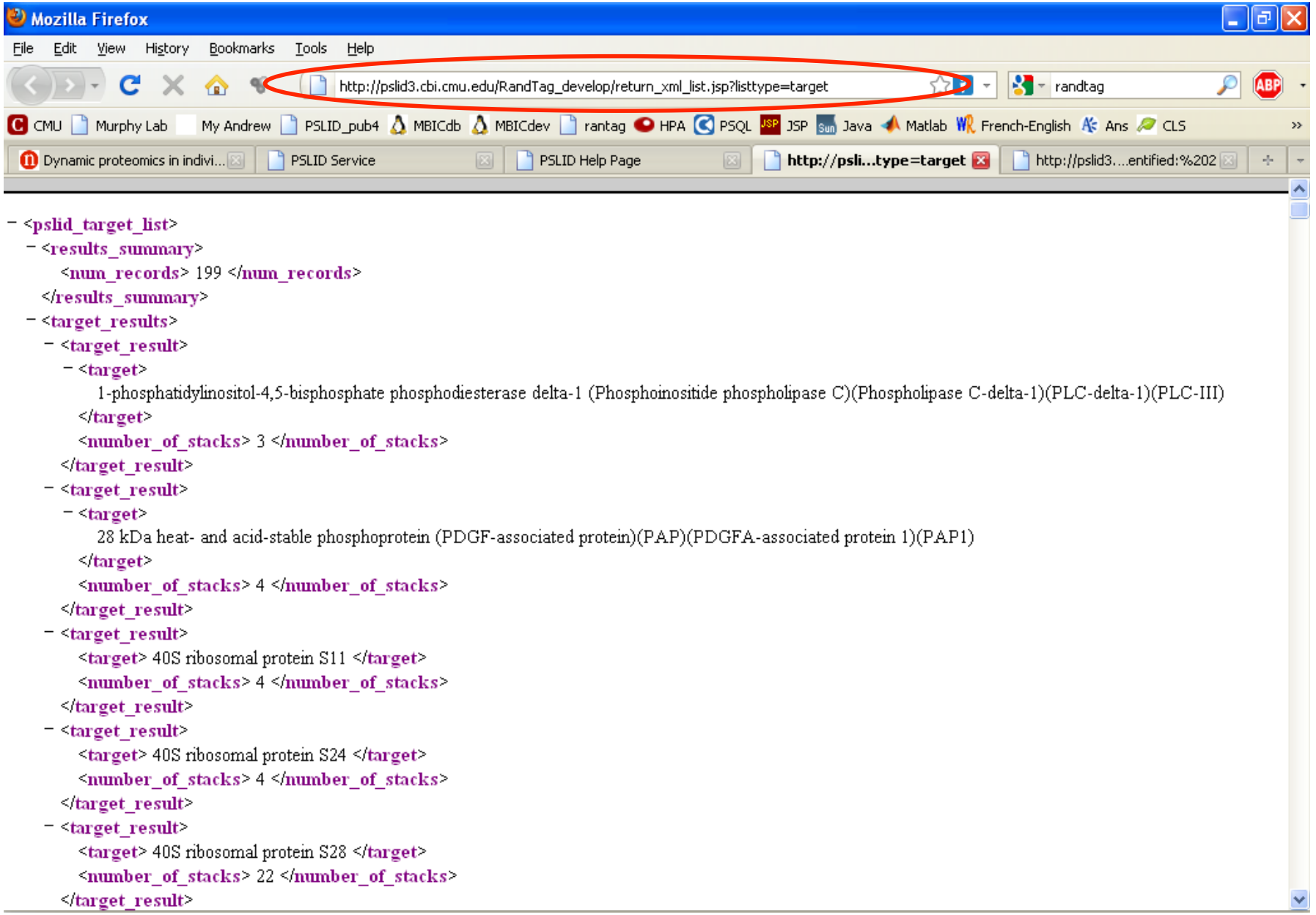
22 images returned (30 images shown) from the query.

View the summary of set **temp3105_128687F45B090C87B76F494C98E049A8**.

Click besides a given image to retrieve similar images in the database.

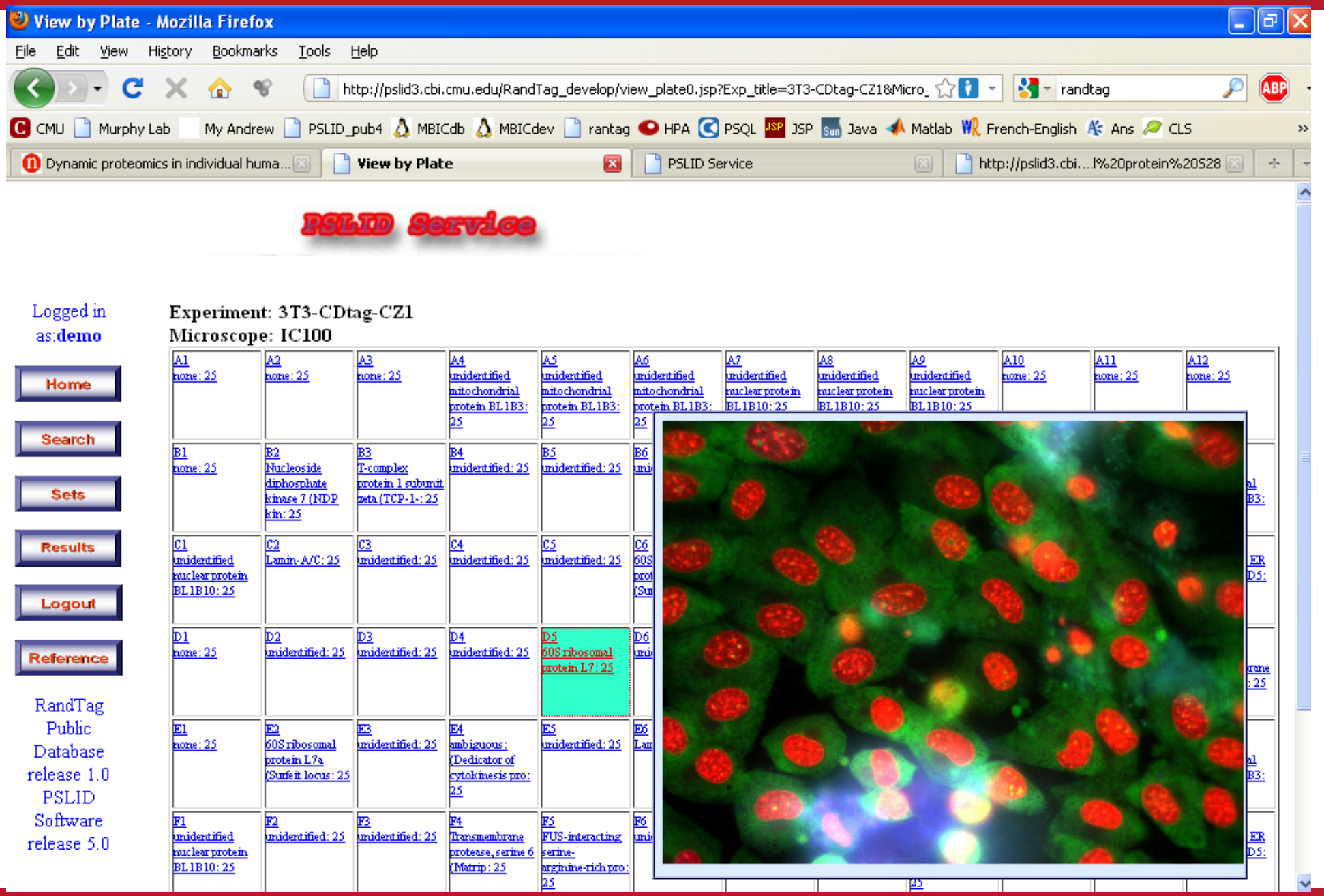
	Image	Cell Name	Organism	Experiment	Protocol	Target	Microscopy & Filter
	Image 214	3T3-CDtag-BM1B9	Mus musculus	3T3-CDtag-BM1	3T3 RandTag Protocol 2	40S ribosomal protein S28	CellLab IC100 Image Cytometer
	Image 232	3T3-CDtag-BM1B9	Mus musculus	3T3-CDtag-BM1	3T3 RandTag Protocol 2	40S ribosomal protein S28	CellLab IC100 Image Cytometer
	Image 235	3T3-CDtag-BM1B9	Mus musculus	3T3-CDtag-BM1	3T3 RandTag Protocol 2	40S ribosomal protein S28	CellLab IC100 Image Cytometer
	Image 238	3T3-CDtag-BM1B9	Mus musculus	3T3-CDtag-BM1	3T3 RandTag Protocol 2	40S ribosomal protein S28	CellLab IC100 Image Cytometer
	Image 244	3T3-CDtag-BM1B9	Mus musculus	3T3-CDtag-BM1	3T3 RandTag Protocol 2	40S ribosomal protein S28	CellLab IC100 Image Cytometer
	Image 247	3T3-CDtag-BM1B9	Mus musculus	3T3-CDtag-BM1	3T3 RandTag Protocol 2	40S ribosomal protein S28	CellLab IC100 Image Cytometer

URL search: by target, returned in XML format

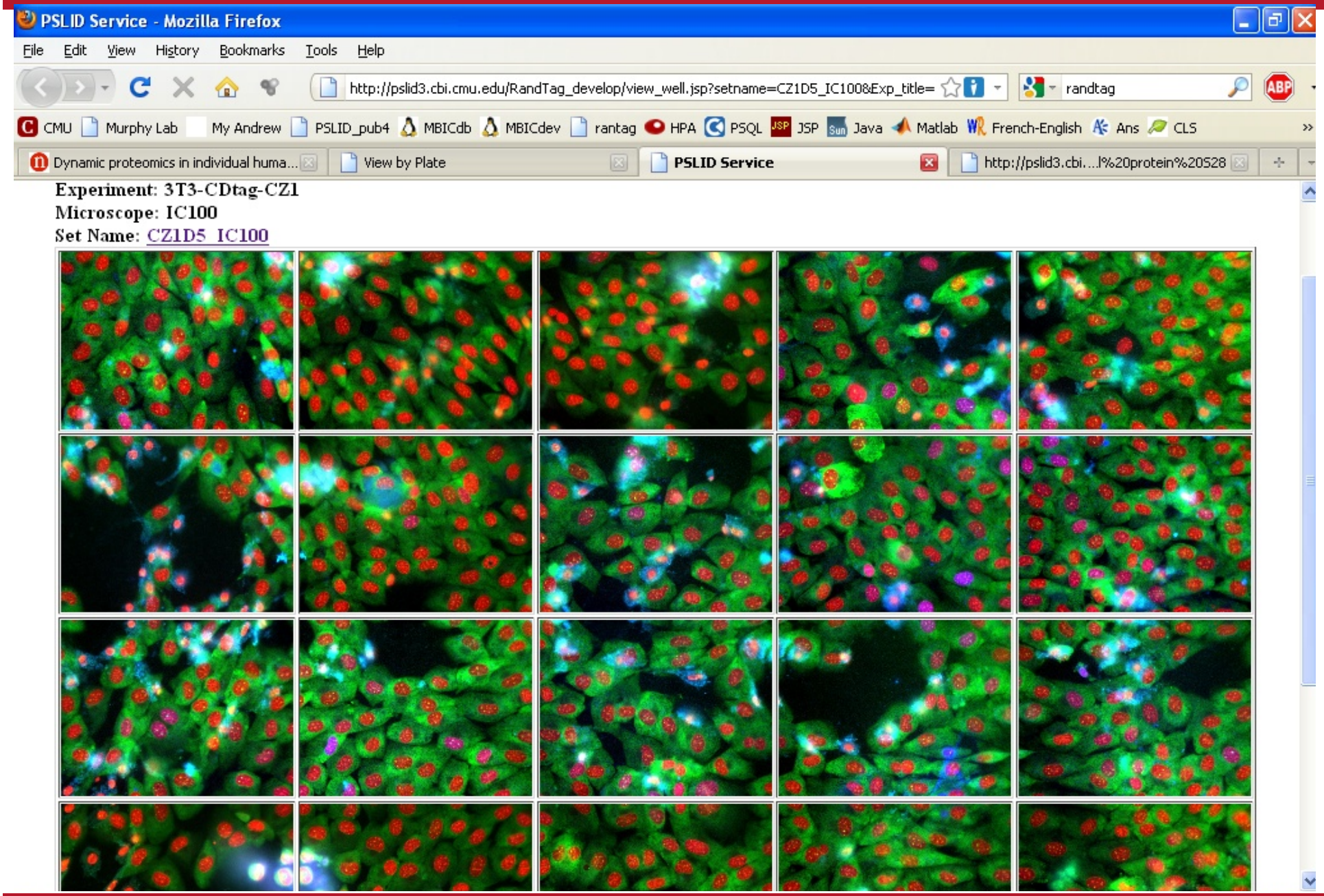



```
- <pslid_target_list>
  - <results_summary>
      <num_records> 199 </num_records>
    </results_summary>
  - <target_results>
    - <target_result>
      - <target>
          1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta-1 (Phosphoinositide phospholipase C)(Phospholipase C-delta-1)(PLC-delta-1)(PLC-III)
        </target>
        <number_of_stacks> 3 </number_of_stacks>
      </target_result>
    - <target_result>
      - <target>
          28 kDa heat- and acid-stable phosphoprotein (PDGF-associated protein)(PAP)(PDGFA-associated protein 1)(PAP1)
        </target>
        <number_of_stacks> 4 </number_of_stacks>
      </target_result>
    - <target_result>
        <target> 40S ribosomal protein S11 </target>
        <number_of_stacks> 4 </number_of_stacks>
      </target_result>
    - <target_result>
        <target> 40S ribosomal protein S24 </target>
        <number_of_stacks> 4 </number_of_stacks>
      </target_result>
    - <target_result>
        <target> 40S ribosomal protein S28 </target>
        <number_of_stacks> 22 </number_of_stacks>
      </target_result>
```

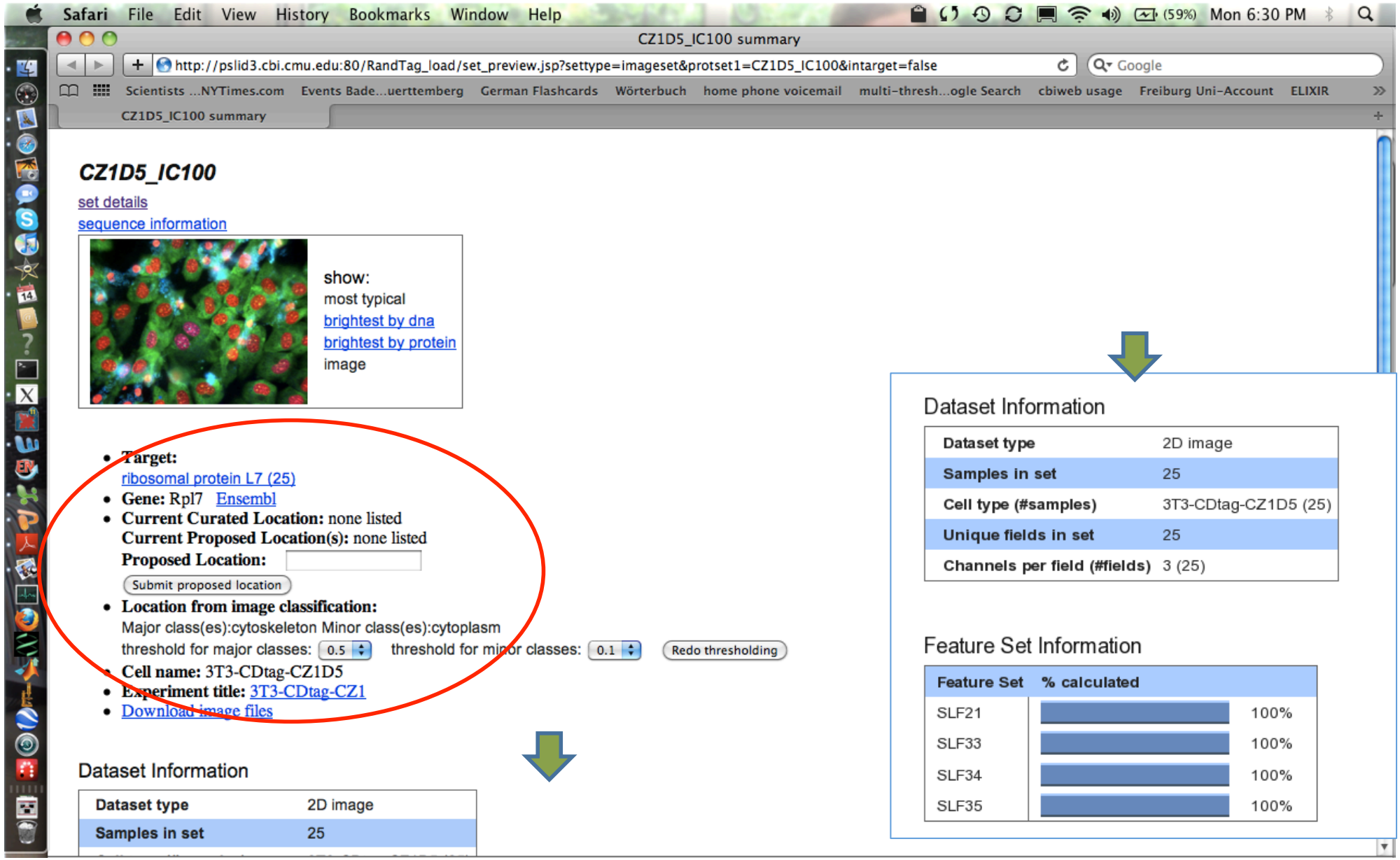
View by plate



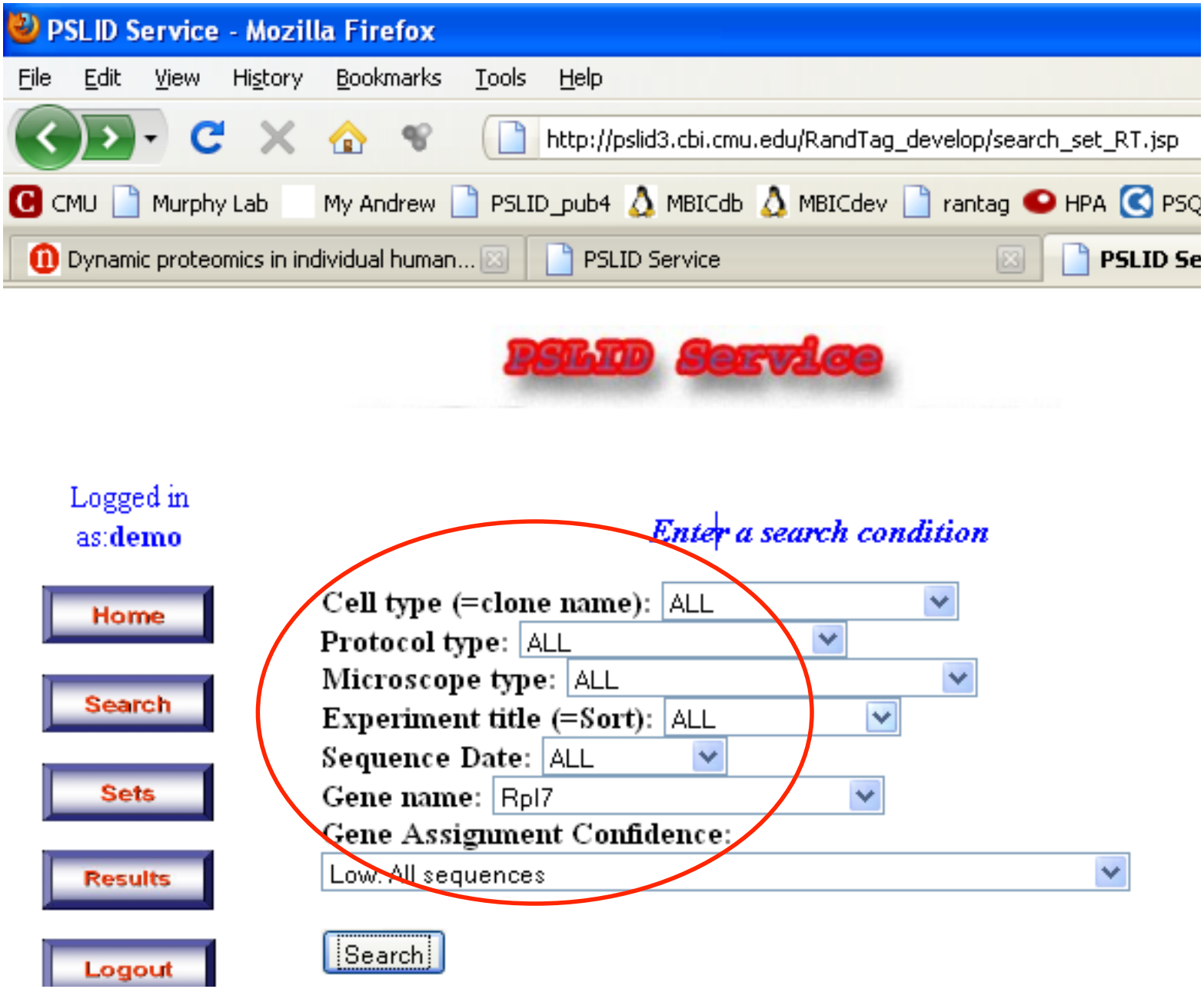
Images of one well



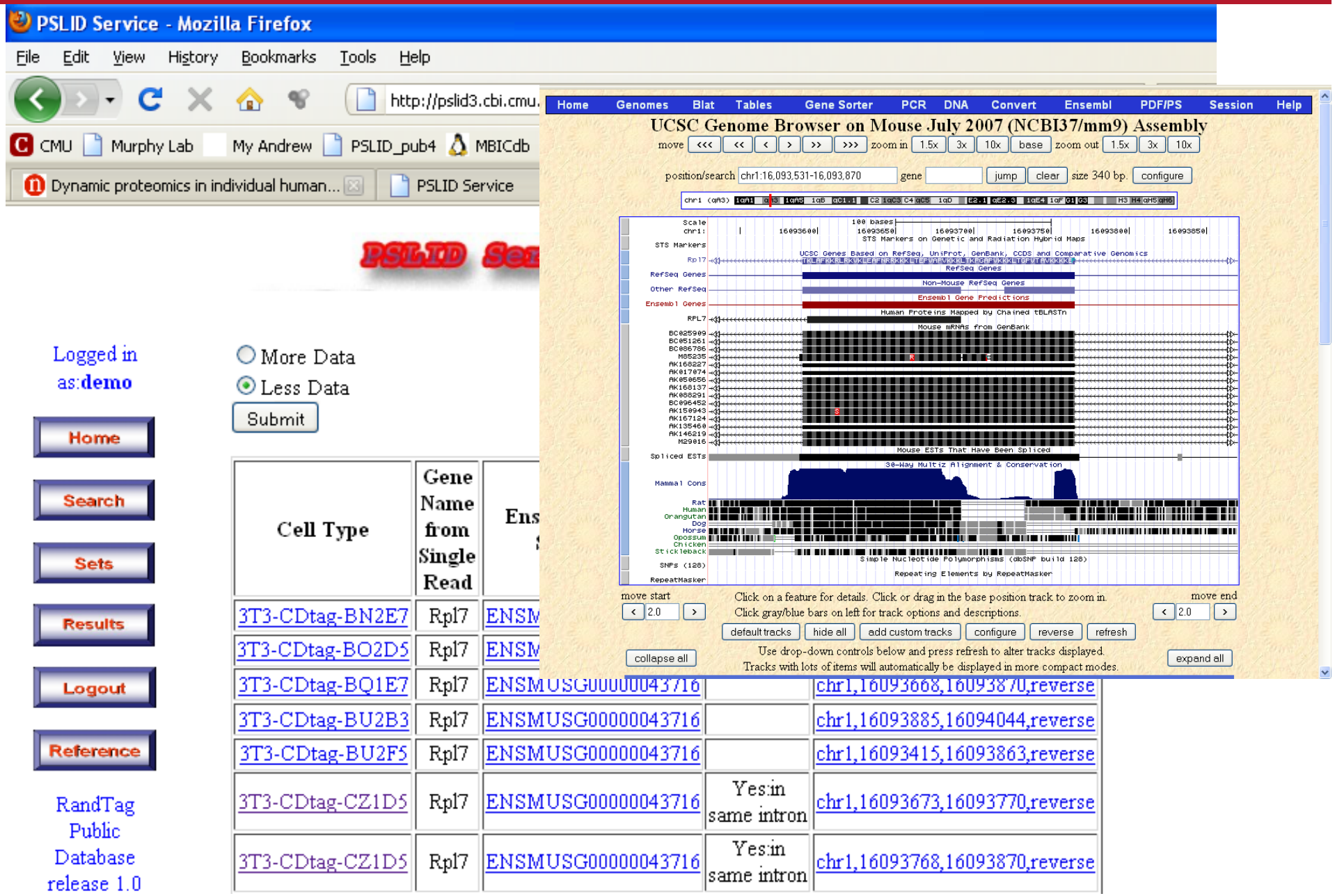
View Set Summary



Multi criterion search



Search by Gene Name



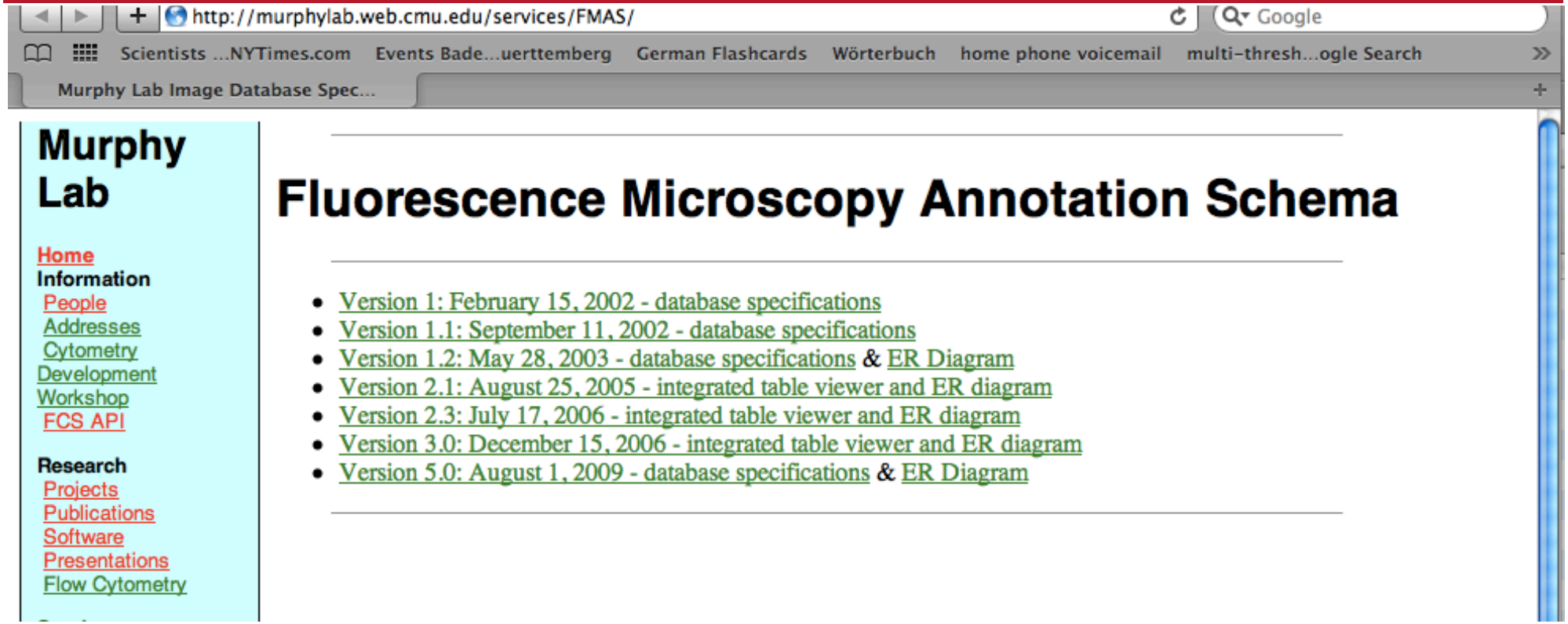
Analysis tools

- Segmentation/feature calculation
- Typical image selection
- Set comparison
- Feature selection
- Classifier training/use
- Stacked/meta classifier training/use
- Clustering to sets
- Generative models training/synthesis

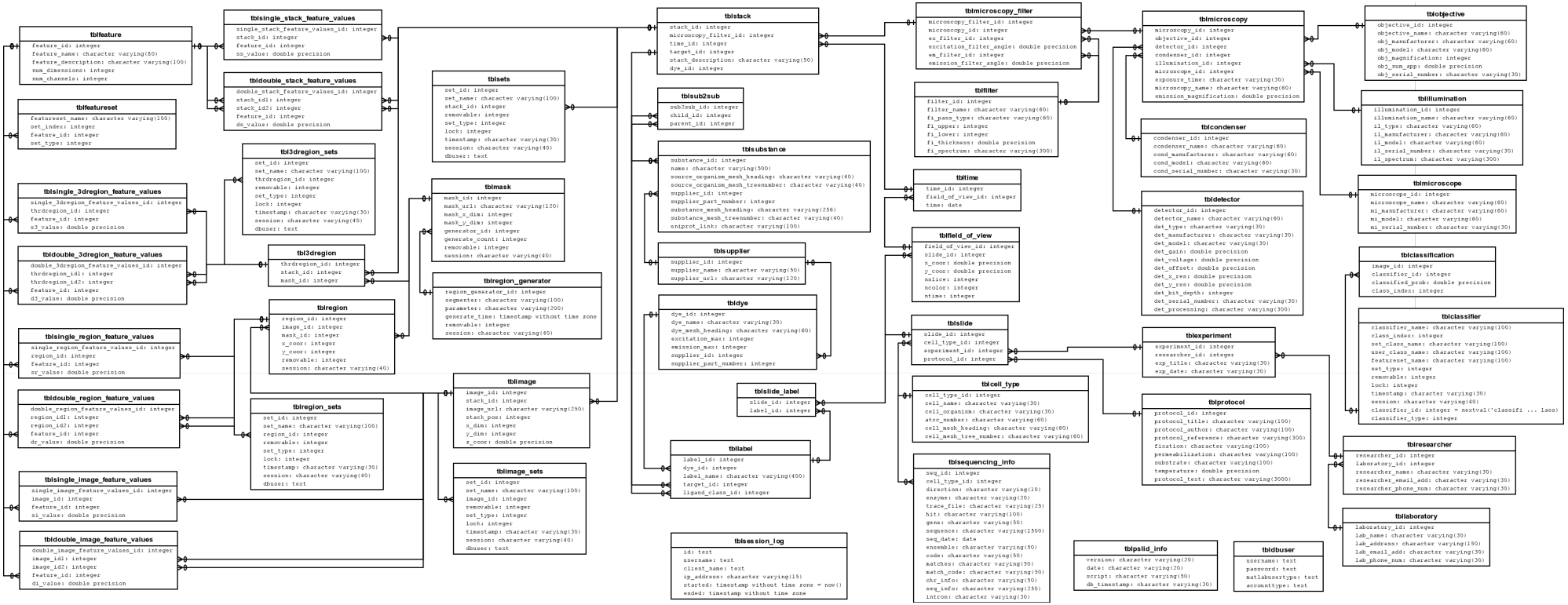
Exporting from PSLID

- For any set, can export
  - Feature file for selected feature set
    - Matlab .mat file
    - CSV
    - Tab delimited
  - Zip file containing all images
    - Original format
    - OME TIFF
      - Images
      - Metadata (OME XML + PSLID custom anotations)
      - Analysis results
    - HDF5 (coming)

PSLID data model



FMAS schema


tblfeature
feature_id: integer
feature_name: character varying(80)
feature_description: character varying(100)
num_dimensions: integer
num_channels: integer
tblfeatureset
featureset_name: character varying(100)
set_index: integer
feature_id: integer
set_type: integer
tblsingle_stack_feature_values
single_stack_feature_values_id: integer
stack_id: integer
feature_id: integer
ss_value: double precision
tbldouble_stack_feature_values
double_stack_feature_values_id: integer
stack_id1: integer
stack_id2: integer
feature_id: integer
ds_value: double precision
tbl3dregion_sets
set_id: integer
set_name: character varying(100)
thrdregion_id: integer
removable: integer
set_type: integer
lock: integer
timestamp: character varying(30)
session: character varying(40)
dbuser: text
tblsingle_3dregion_feature_values
single_3dregion_feature_values_id: integer
thrdregion_id: integer
feature_id: integer
s3_value: double precision
tbldouble_3dregion_feature_values
double_3dregion_feature_values_id: integer
thrdregion_id1: integer
thrdregion_id2: integer
feature_id: integer
d3_value: double precision
tbl3dregion
thrdregion_id: integer
stack_id: integer
mask_id: integer
tblsets
set_id: integer
set_name: character varying(100)
stack_id: integer
removable: integer
set_type: integer
lock: integer
timestamp: character varying(30)
session: character varying(40)
dbuser: text
tblmask
mask_id: integer
mask_url: character varying(120)
mask_x_dim: integer
mask_y_dim: integer
generator_id: integer
generate_count: integer
removable: integer
session: character varying(40)
tblregion_generator
region_generator_id: integer
segmenter: character varying(100)
parameter: character varying(200)
generate_time: timestamp without time zone
removable: integer
session: character varying(40)
tblregion
region_id: integer
image_id: integer
mask_id: integer
x_coor: integer
y_coor: integer
removable: integer
session: character varying(40)
tblsingle_region_feature_values
single_region_feature_values_id: integer
region_id: integer
feature_id: integer
sr_value: double precision
tbldouble_region_feature_values
double_region_feature_values_id: integer
region_id1: integer
region_id2: integer
feature_id: integer
dr_value: double precision
tblregion_sets
set_id: integer
set_name: character varying(100)
region_id: integer
removable: integer
set_type: integer
lock: integer
timestamp: character varying(30)
session: character varying(40)
dbuser: text
tblsingle_image_feature_values
single_image_feature_values_id: integer
image_id: integer
feature_id: integer
si_value: double precision
tbldouble_image_feature_values
double_image_feature_values_id: integer
image_id1: integer
image_id2: integer
feature_id: integer
di_value: double precision
tblimage
image_id: integer
stack_id: integer
image_url: character varying(250)
stack_pos: integer
x_dim: integer
y_dim: integer
z_coor: double precision
tblimage_sets
set_id: integer
set_name: character varying(100)
image_id: integer
removable: integer
set_type: integer
lock: integer
timestamp: character varying(30)
session: character varying(40)
dbuser: text
tblstack
stack_id: integer
microscopy_filter_id: integer
time_id: integer
target_id: integer
stack_description: character varying(50)
dye_id: integer
tblsub2sub
ssb2sub_id: integer
child_id: integer
parent_id: integer
tblsubstance
substance_id: integer
name: character varying(500)
source_organism_mesh_heading: character varying(40)
source_organism_mesh_treenumber: character varying(40)
supplier_id: integer
supplier_part_number: integer
substance_mesh_heading: character varying(256)
substance_mesh_treenumber: character varying(40)
uniprot_link: character varying(100)
tblsupplier
supplier_id: integer
supplier_name: character varying(50)
supplier_url: character varying(120)
tbldye
dye_id: integer
dye_name: character varying(30)
dye_mesh_heading: character varying(60)
excitation_max: integer
emission_max: integer
supplier_id: integer
supplier_part_number: integer
tblslide_label
slide_id: integer
label_id: integer
tbllabel
label_id: integer
dye_id: integer
label_name: character varying(400)
target_id: integer
ligand_class_id: integer
tblsession_log
id: text
username: text
client_name: text
ip_address: character varying(15)
started: timestamp without time zone = now()
ended: timestamp without time zone
tblmicroscopy_filter
microscopy_filter_id: integer
microscopy_id: integer
ex_filter_id: integer
excitation_filter_angle: double precision
em_filter_id: integer
emission_filter_angle: double precision
tblfilter
filter_id: integer
filter_name: character varying(60)
fi_pass_type: character varying(60)
fi_upper: integer
fi_lower: integer
fi_thickness: double precision
fi_spectrum: character varying(300)
tbltime
time_id: integer
field_of_view_id: integer
time: date
tblfield_of_view
field_of_view_id: integer
slide_id: integer
x_coor: double precision
y_coor: double precision
nslice: integer
ncolor: integer
ntime: integer
tblslide
slide_id: integer
cell_type_id: integer
experiment_id: integer
protocol_id: integer
tblcell_type
cell_type_id: integer
cell_name: character varying(30)
cell_organism: character varying(30)
atcc_number: character varying(60)
cell_mesh_heading: character varying(60)
cell_mesh_tree_number: character varying(60)
tblsequencing_info
seq_id: integer
cell_type_id: integer
direction: character varying(20)
enzyme: character varying(20)
trace_file: character varying(25)
hit: character varying(100)
gene: character varying(50)
sequence: character varying(1500)
seq_date: date
ensemble: character varying(50)
code: character varying(50)
matches: character varying(50)
match_code: character varying(50)
chr_info: character varying(50)
seq_info: character varying(250)
introns: character varying(30)
tblmicroscopy
microscopy_id: integer
objective_id: integer
detector_id: integer
condenser_id: integer
illumination_id: integer
microscope_id: integer
exposure_time: character varying(30)
microscopy_name: character varying(60)
emission_magnification: double precision
tblcondenser
condenser_id: integer
condenser_name: character varying(60)
cond_manufacturer: character varying(60)
cond_model: character varying(60)
cond_serial_number: character varying(30)
tbldetector
detector_id: integer
detector_name: character varying(60)
det_type: character varying(30)
det_manufacturer: character varying(30)
det_model: character varying(30)
det_gain: double precision
det_voltage: double precision
det_offset: double precision
det_x_res: double precision
det_y_res: double precision
det_bit_depth: integer
det_serial_number: character varying(30)
det_processing: character varying(300)
tblexperiment
experiment_id: integer
researcher_id: integer
exp_title: character varying(30)
exp_date: character varying(30)
tblprotocol
protocol_id: integer
protocol_title: character varying(100)
protocol_author: character varying(100)
protocol_reference: character varying(300)
fixation: character varying(100)
permeabilization: character varying(100)
substrate: character varying(100)
temperature: double precision
protocol_text: character varying(3000)
tblpslid_info
version: character varying(20)
date: character varying(20)
script: character varying(50)
db_timestamp: character varying(30)
tbldbuser
username: text
password: text
matlabusertype: text
accounttype: text
tblobjective
objective_id: integer
objective_name: character varying(60)
obj_manufacturer: character varying(60)
obj_model: character varying(60)
obj_magnification: integer
obj_num_app: double precision
obj_serial_number: character varying(30)
tblillumination
illumination_id: integer
illumination_name: character varying(60)
il_type: character varying(60)
il_manufacturer: character varying(60)
il_model: character varying(60)
il_serial_number: character varying(30)
il_spectrum: character varying(300)
tblmicroscope
microscope_id: integer
microscope_name: character varying(60)
mi_manufacturer: character varying(60)
mi_model: character varying(60)
mi_serial_number: character varying(30)
tblclassification
image_id: integer
classifier_id: integer
classified_prob: double precision
class_index: integer
tblclassifier
classifier_name: character varying(100)
class_index: integer
set_class_name: character varying(100)
user_class_name: character varying(100)
featureset_name: character varying(100)
set_type: integer
removable: integer
lock: integer
timestamp: character varying(30)
session: character varying(40)
classifier_id: integer = nextval('classifi ... lass)
classifier_type: integer
tblresearcher
researcher_id: integer
laboratory_id: integer
researcher_name: character varying(30)
researcher_email_add: character varying(30)
researcher_phone_num: character varying(30)
tbllaboratory
laboratory_id: integer
lab_name: character varying(30)
lab_address: character varying(150)
lab_email_add: character varying(30)
lab_phone_num: character varying(30)

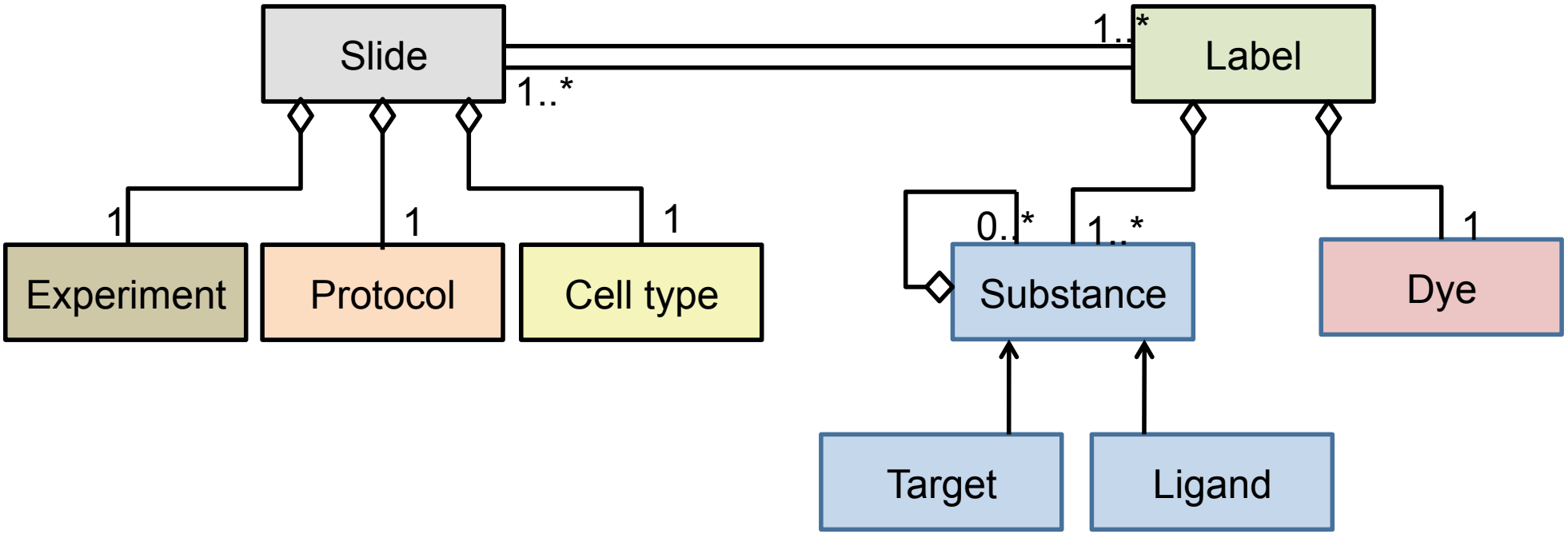
Importing into OMERO

- OME TIFF with PSLID custom annotations
  - Works with current OMERO.importer
  - Can’t load/read custom annotations
  - should be fixed in OMERO 4.2
- Currently require additional import into PSLID server to capture the rest of the annotations

Adding PSLID tags to OME-XML

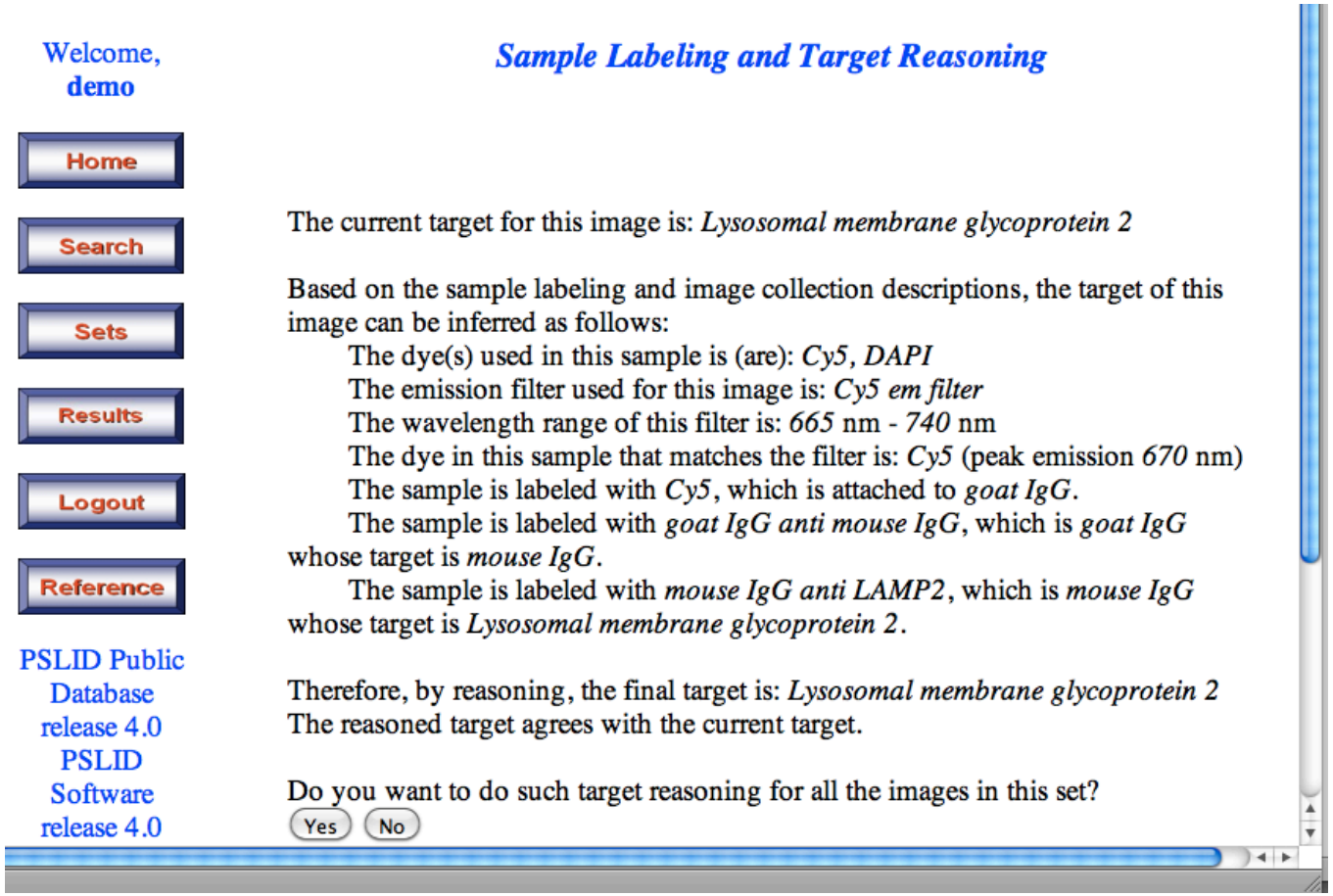
- Description of cell lines, reagents and probes in sample
- Features, classifications, and other results from analysis

Sample Description



Target reasoning

- Annotations specify
  - excitation wavelength and emission filters for each image
  - labels (substances, dyes) present in each sample
- Code reasons out which target substance is visualized in each image



Welcome,
demo

Home

Search

Sets

Results

Logout

Reference

PSLID Public Database release 4.0

PSLID Software release 4.0

Sample Labeling and Target Reasoning

The current target for this image is: *Lysosomal membrane glycoprotein 2*

Based on the sample labeling and image collection descriptions, the target of this image can be inferred as follows:

The dye(s) used in this sample is (are): *Cy5*, *DAPI*
The emission filter used for this image is: *Cy5 em filter*
The wavelength range of this filter is: *665* nm - *740* nm
The dye in this sample that matches the filter is: *Cy5* (peak emission *670* nm)
The sample is labeled with *Cy5*, which is attached to *goat IgG*.
The sample is labeled with *goat IgG anti mouse IgG*, which is *goat IgG* whose target is *mouse IgG*.
The sample is labeled with *mouse IgG anti LAMP2*, which is *mouse IgG* whose target is *Lysosomal membrane glycoprotein 2*.

Therefore, by reasoning, the final target is: *Lysosomal membrane glycoprotein 2*
The reasoned target agrees with the current target.

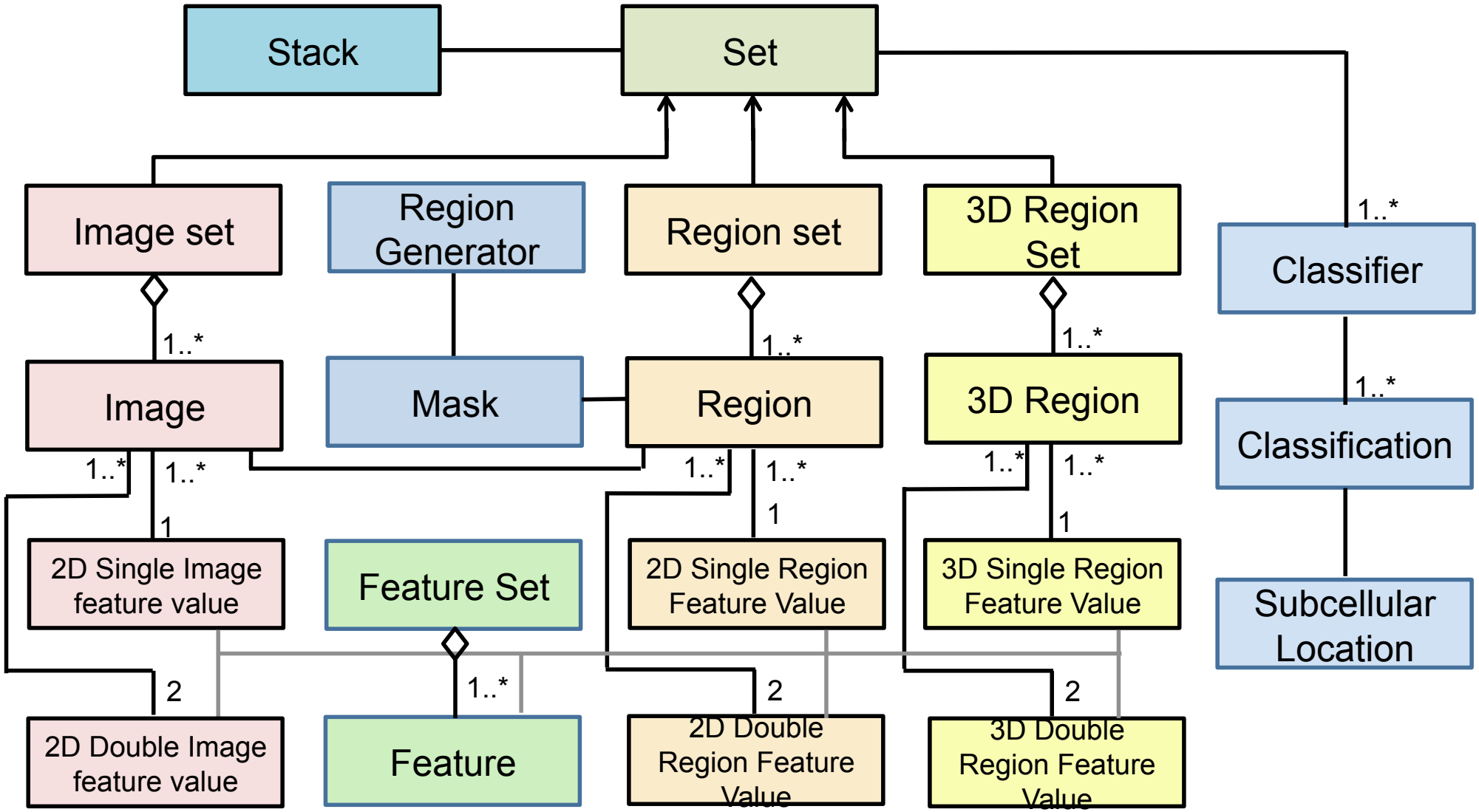
Do you want to do such target reasoning for all the images in this set?

Yes No

Target reasoning

- Permits identification of generalizations and specializations
  - Can search for “tubulin” and find all isoforms
  - Can search for “tubulin-alpha” and be asked “No matches, do you want to see images of tubulin-beta?”
- Permits automatically find differences in reagents and associate them with differences in results

PSLID analysis results metadata



PSLID client and OMERO server

- PSLID client can connect to OMERO server to select images and create a PSLID set from them
- Currently: duplicated in PSLID server
- Planned: replace “plumbing” in PSLID server with calls to OMERO server
  - Feasibility tests completed

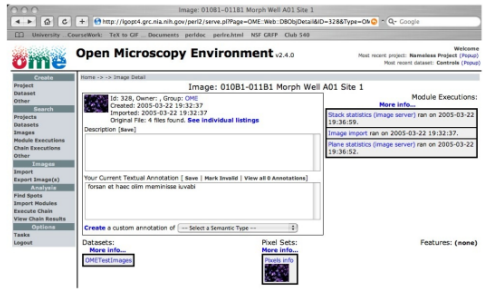
PSLID-OMERO

OMERO.client

-Data management
-Browse images
-Measurement
-**Interact with PSLID**

Upload/retrieve

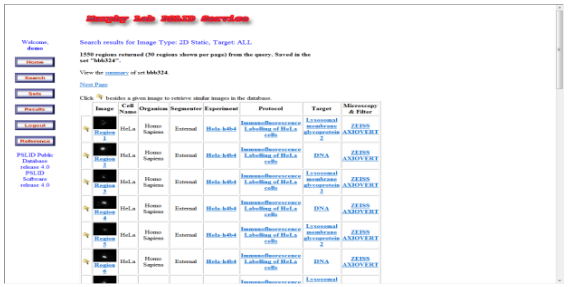
OMERO.server



Upload/retrieve

Invoke PSLID

PSLID client/server

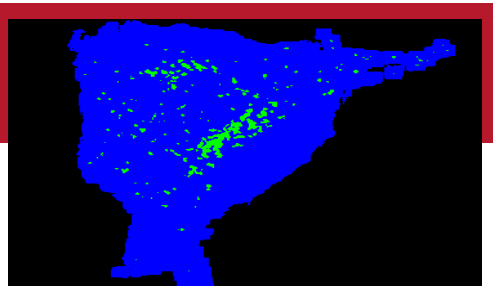


PUnMix: Pattern unmixing

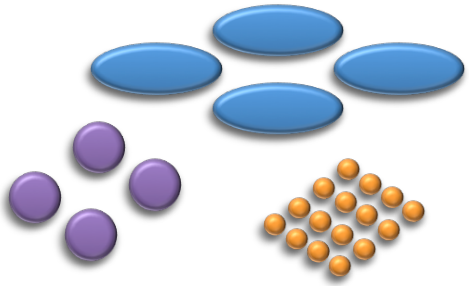
- Many proteins (or other macromolecules) may be found in more than one organelle
- Features “see” each combination of organelles as a new pattern
- Can we “unmix” such mixed patterns?

- Assume that we have markers that are found in only one subcellular location (fundamental pattern)
- Assume that each fundamental pattern can be represented by some combination of distinct object types (e.g., 10% small round objects and 90% long skinny objects)
- Assume that a mixed pattern is formed by adding to gether the objects from fundamental patterns
- Then for each unknown image we can calculate the fraction of fluorescence in each object type, and then estimate how much of each fundamental pattern must be present

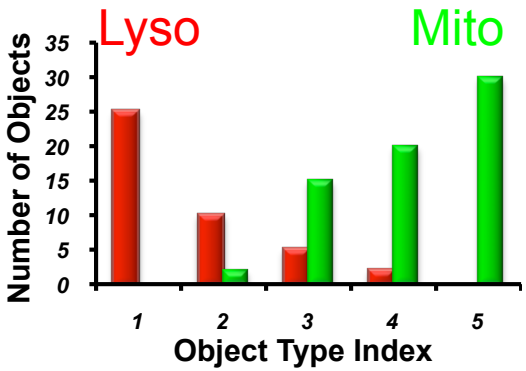
Find objects in each image



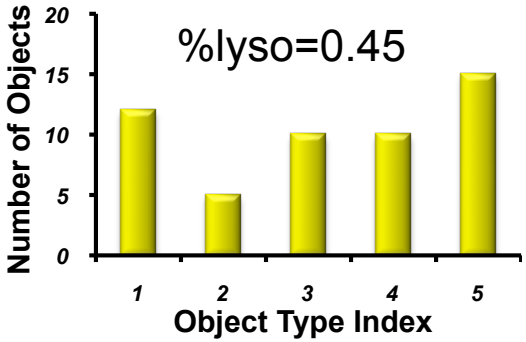
Learn object types from pure samples



Learn model for distributions of object types in pur samples

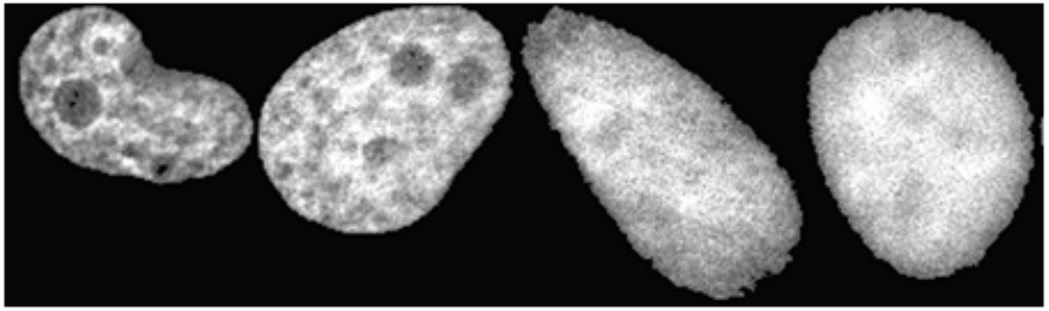


Determine pattern fractions for mixed images

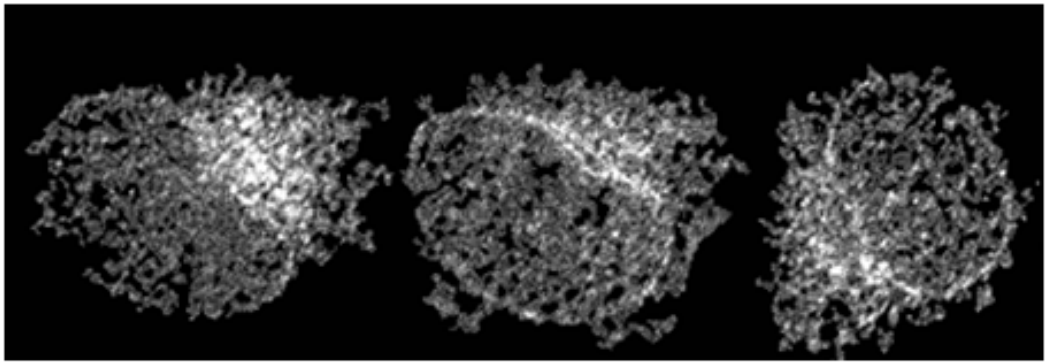


Examples of Object Types

Type A



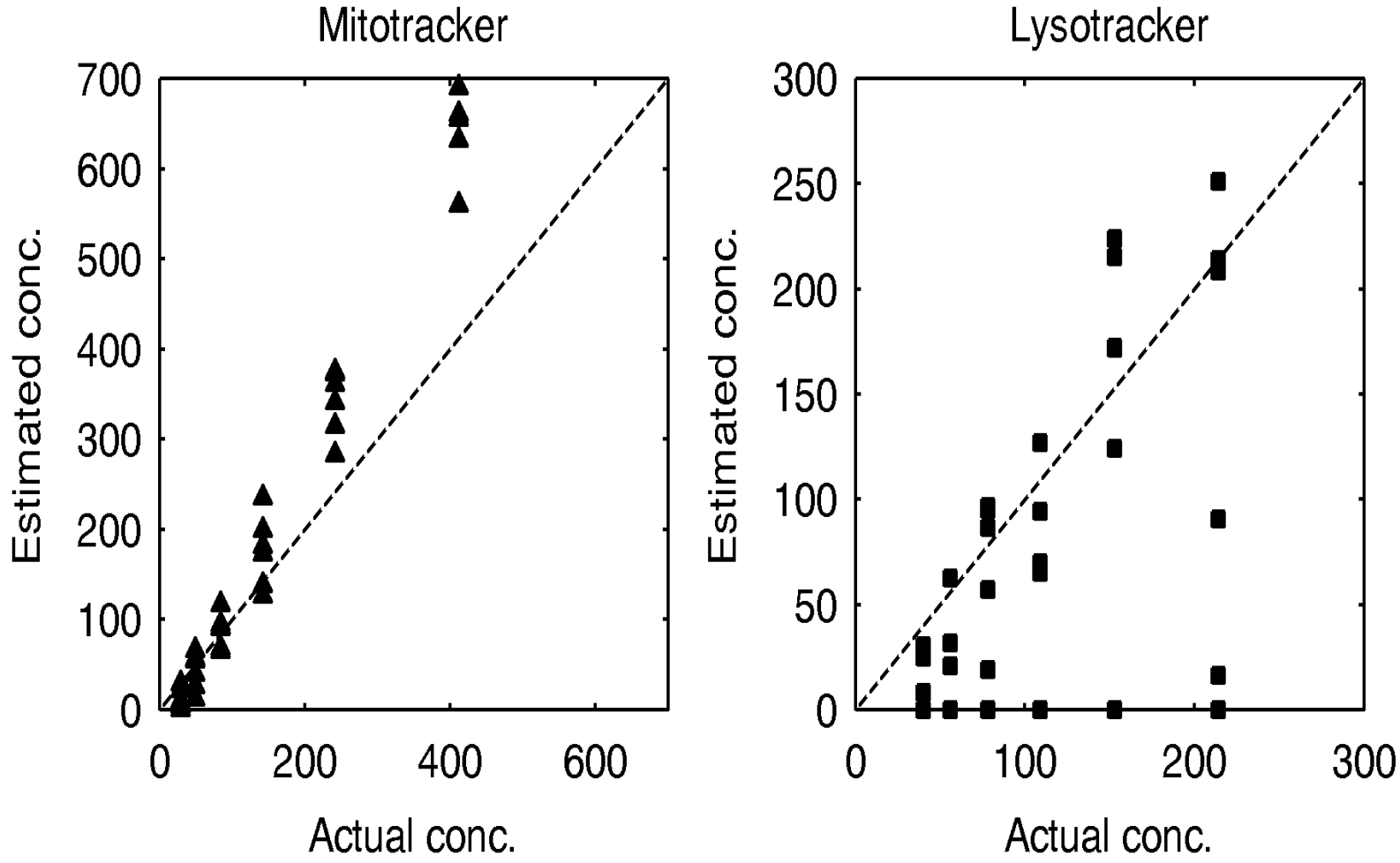
Type B



Type C

Type D

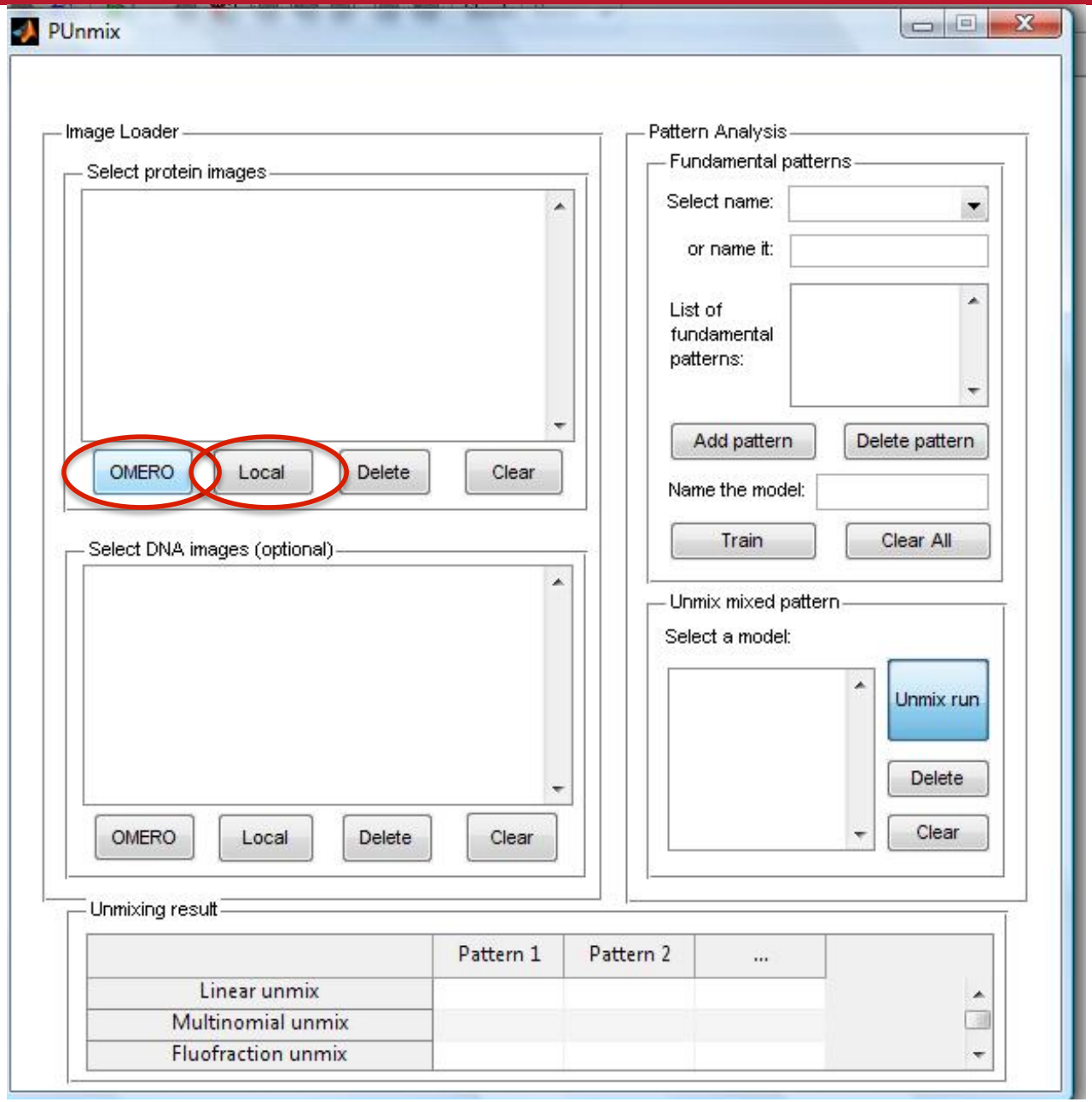
Supervised unmixing results



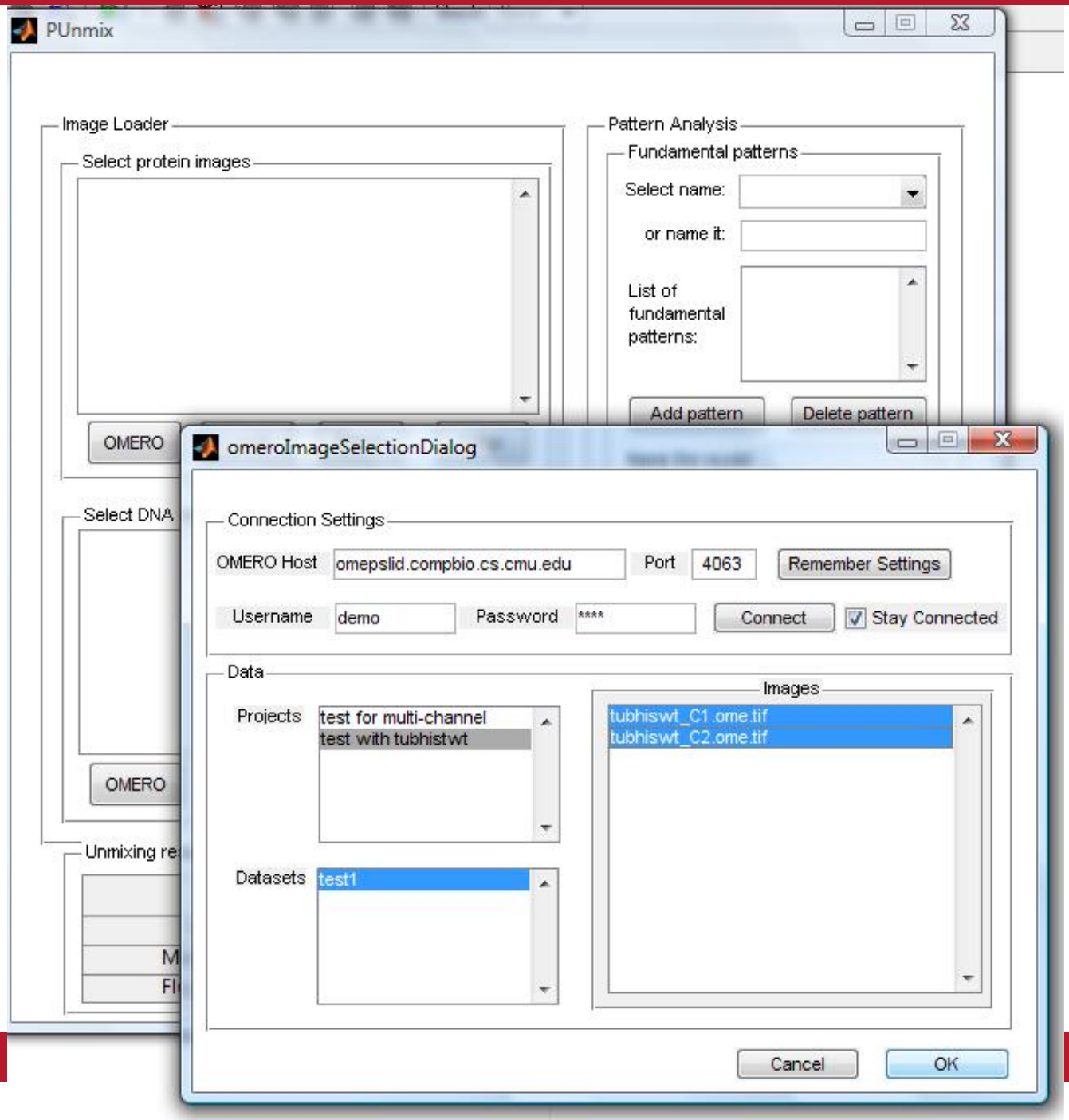
PUnmix and OMERO

- New PUnmix release July 11, 2010
- Versions
  - Matlab source
  - Compiled matlab for MacOS, Windows, Linux
- Supports training on or unmixing images from local disk or multiple OMERO servers

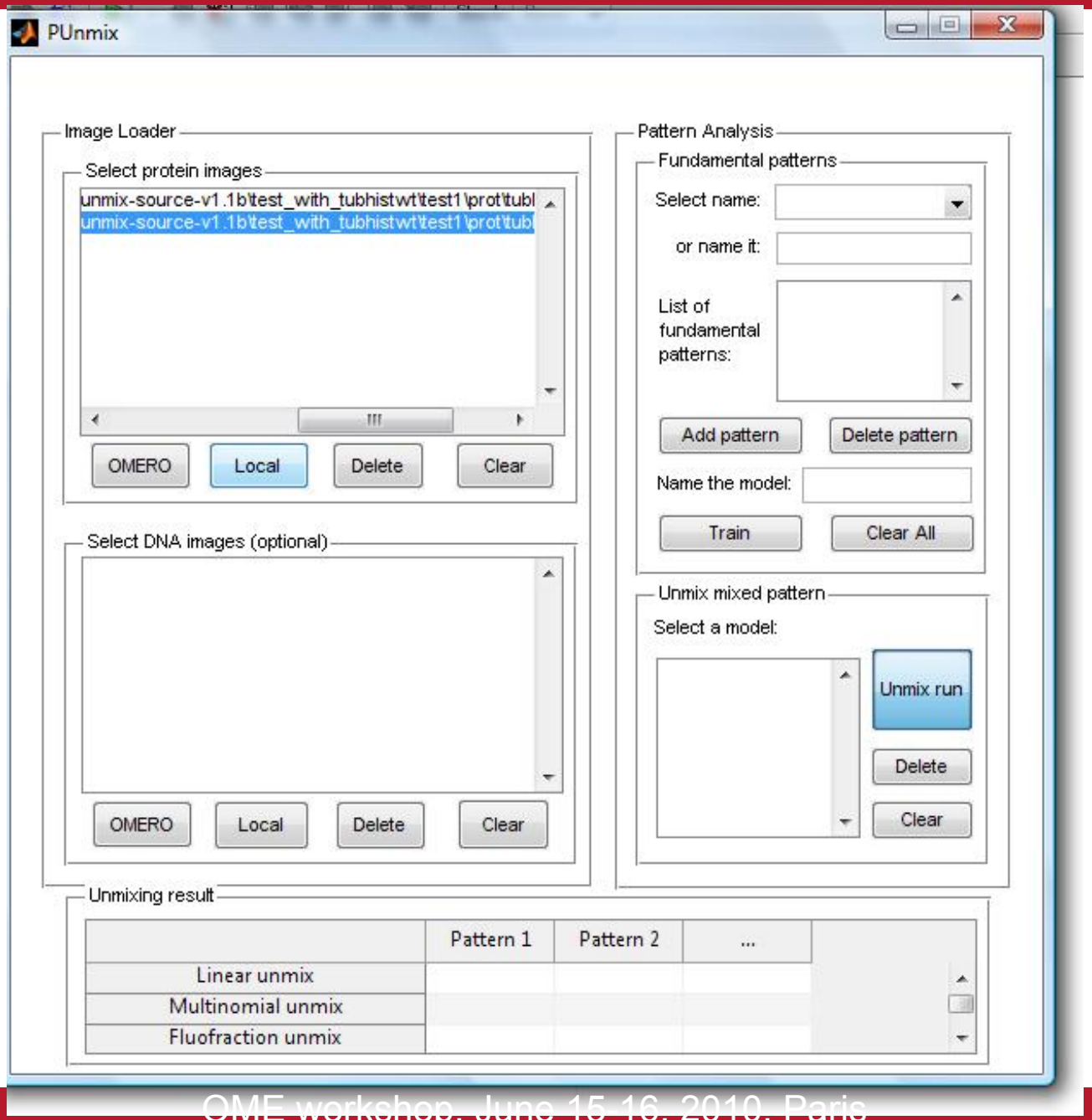
PUnmix GUI


PUnmix
Image Loader
Select protein images
OMERO
Local
Delete
Clear
Select DNA images (optional)
OMERO
Local
Delete
Clear
Pattern Analysis
Fundamental patterns
Select name:
or name it:
List of fundamental patterns:
Add pattern
Delete pattern
Name the model:
Train
Clear All
Unmix mixed pattern
Select a model:
Unmix run
Delete
Clear
Unmixing result
Pattern 1
Pattern 2
...
Linear unmix
Multinomial unmix
Fluofraction unmix

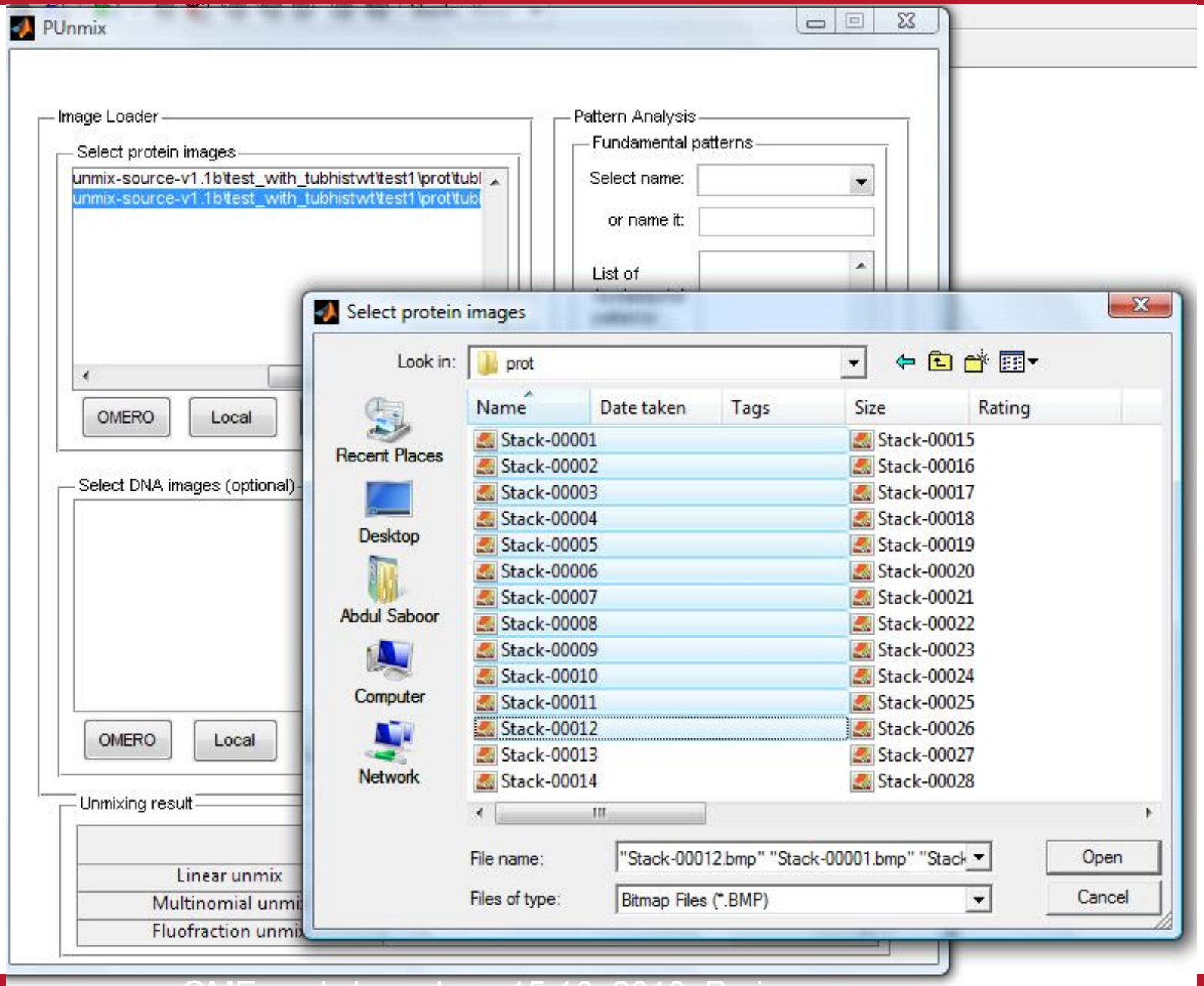
PUnmix GUI


PUnmix
Image Loader
Select protein images
OMERO
Select DNA
OMERO
Unmixing re
Pattern Analysis
Fundamental patterns
Select name:
or name it:
List of fundamental patterns:
Add pattern
Delete pattern
omeroImageSelectionDialog
Connection Settings
OMERO Host
omepslid.compbio.cs.cmu.edu
Port
4063
Remember Settings
Username
demo
Password
****
Connect
Stay Connected
Data
Projects
test for multi-channel
test with tubhistwt
Datasets
test1
Images
tubhiswt_C1.ome.tif
tubhiswt_C2.ome.tif
Cancel
OK

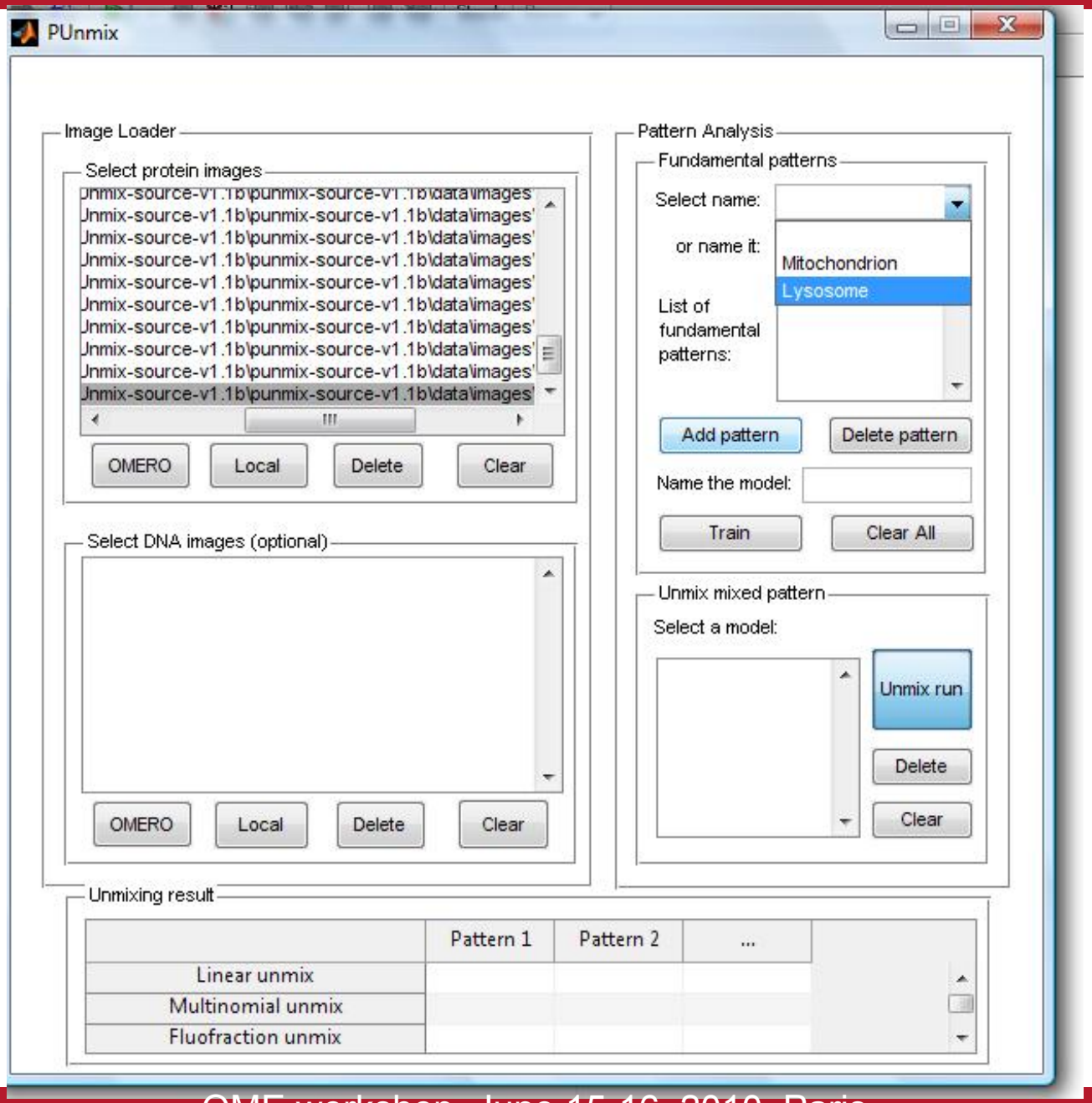
PUnmix after import from OMERO



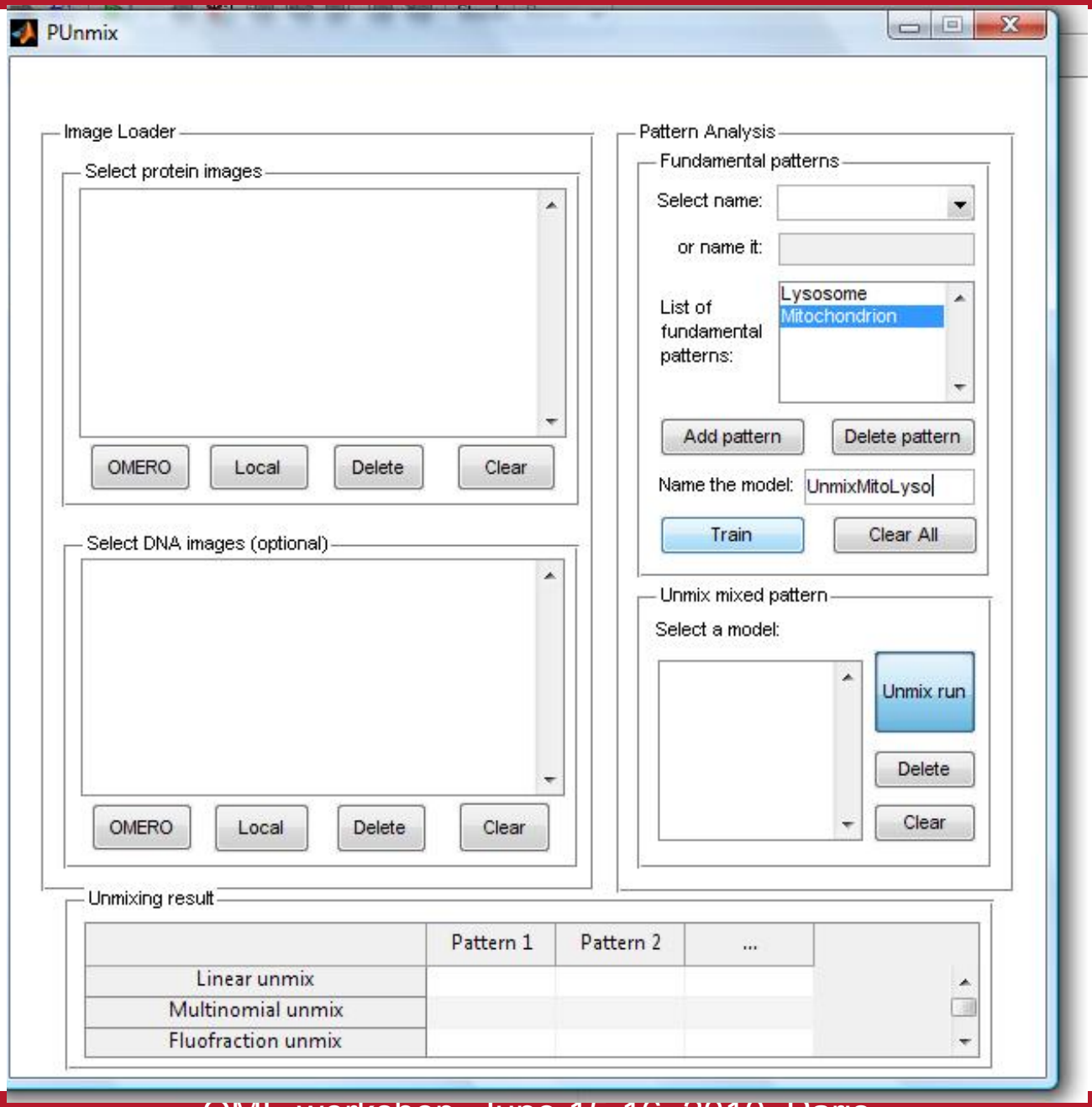
PUnmix adding local images



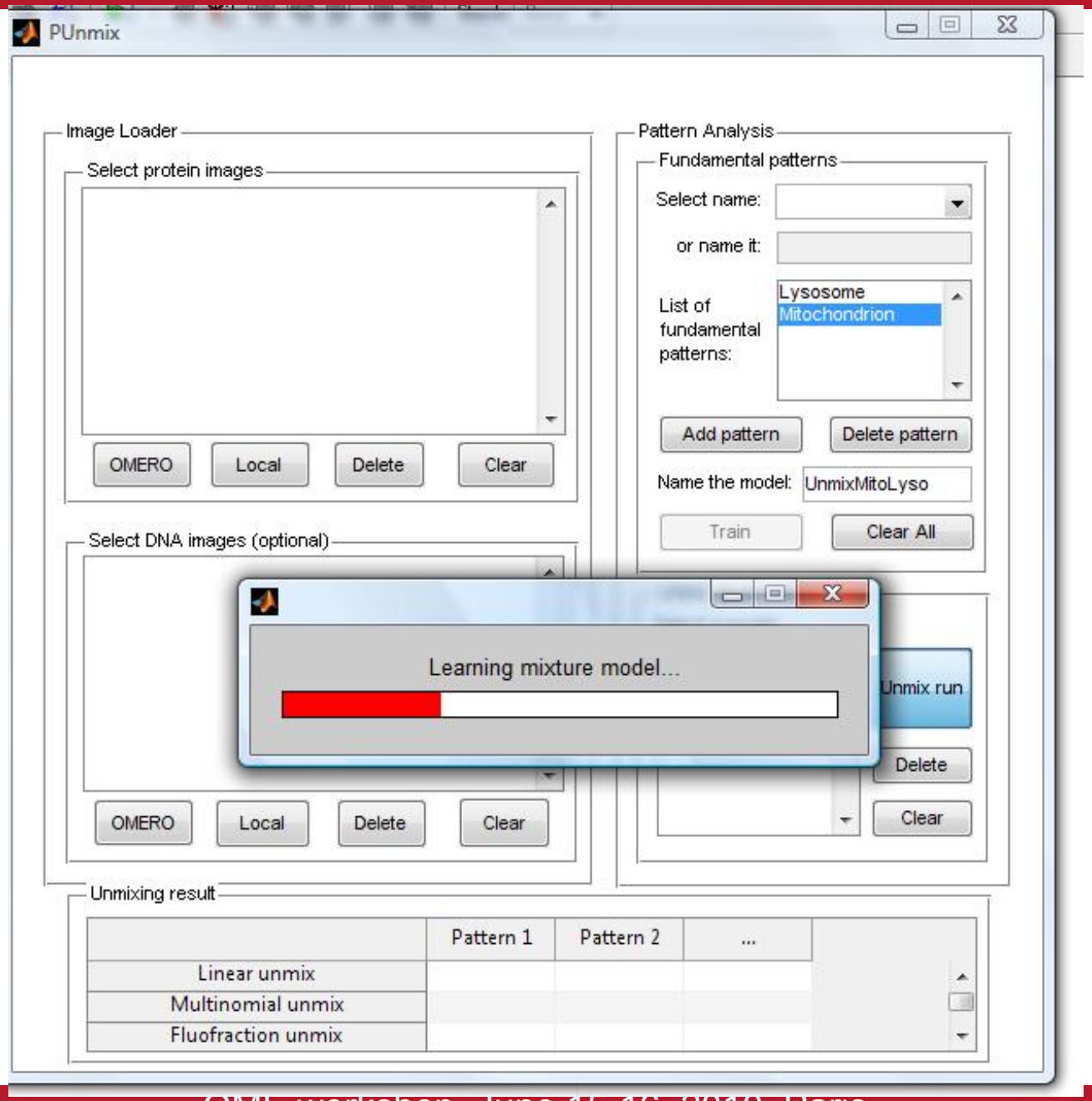
PUnmix training mixing model



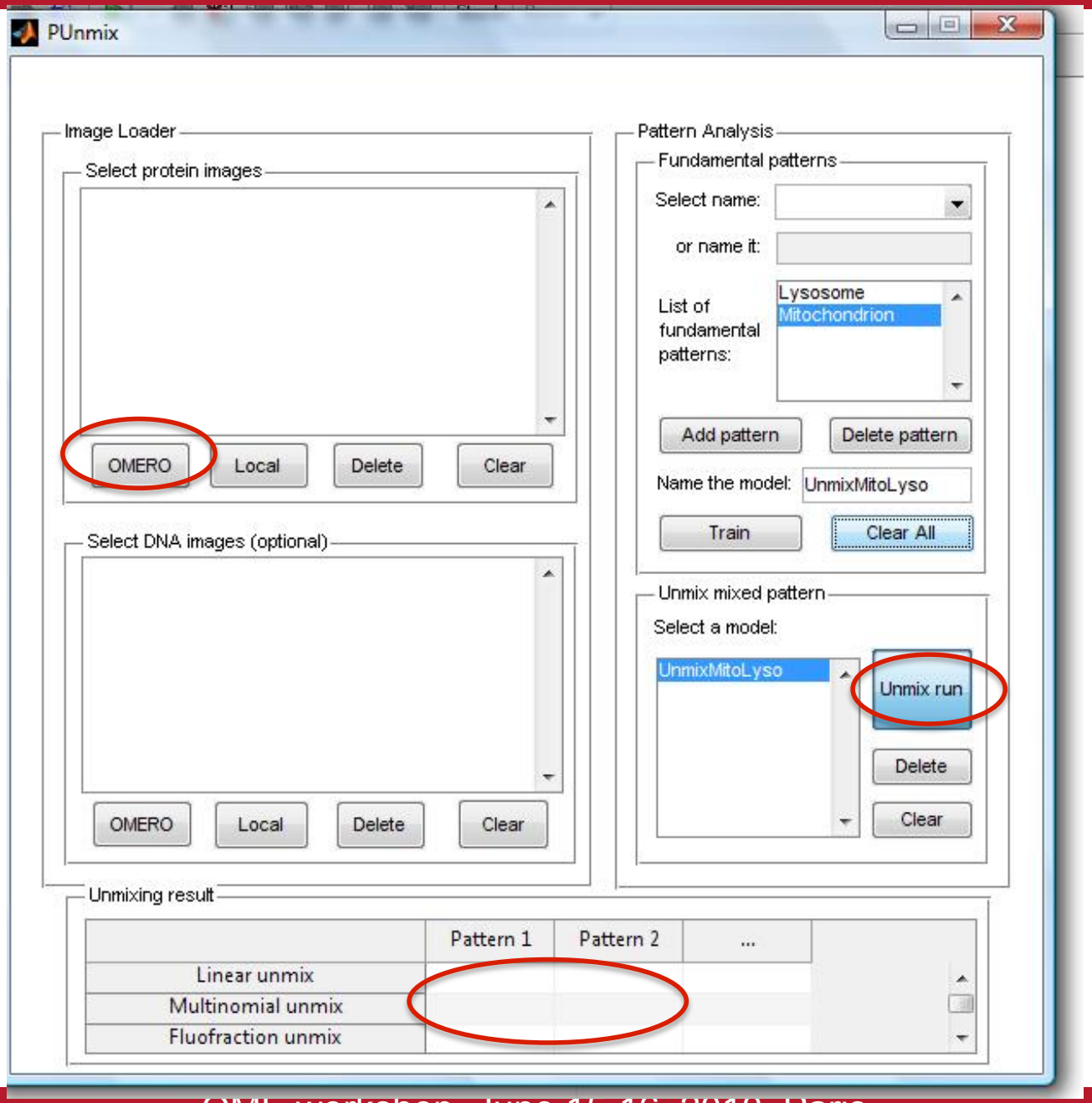
PUnmix training mixing model



PUnmix training mixing model



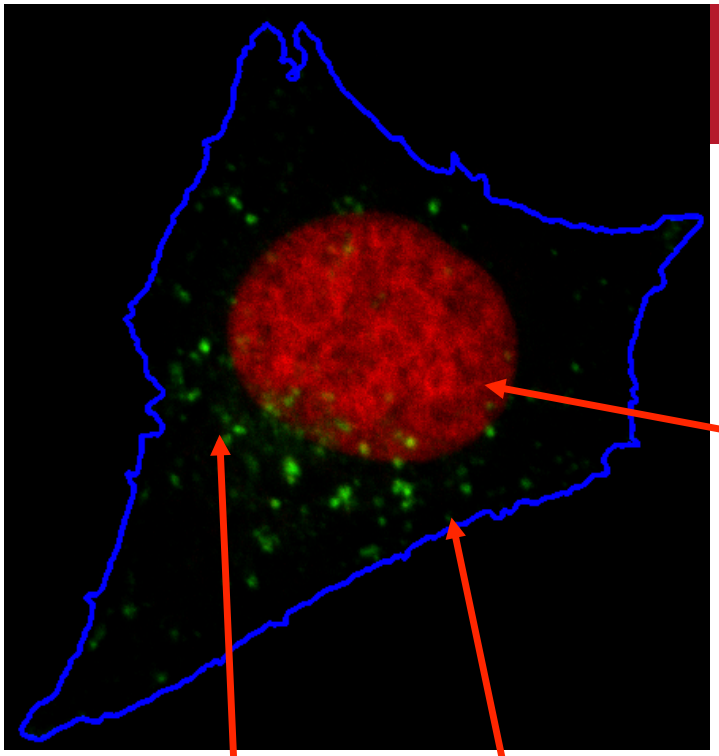
PUnmix training mixing model



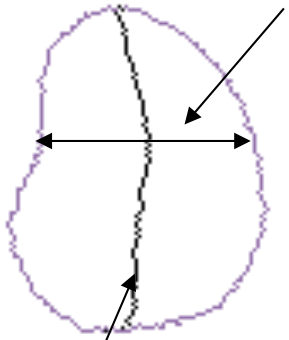
SLML Suite

- SLML suite (Subcellular Location Markup Language) which can learn generative models of subcellular patterns from images and can synthesize new images from them

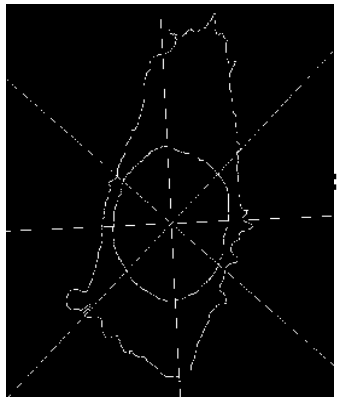
Generative Cell Models



Nucleus



Cell membrane

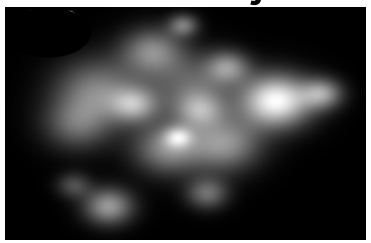


$$\frac{d_1 + d_2}{d_2}$$

Model parameters

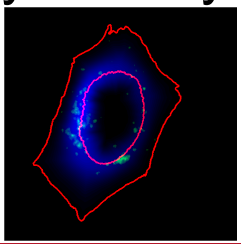
Protein distribution

Gaussian objects



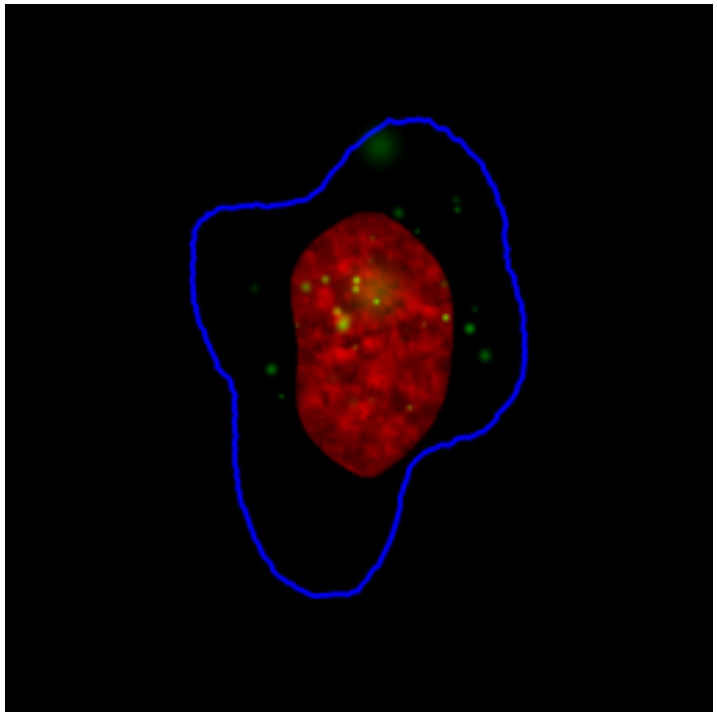
+

Probability Density Function

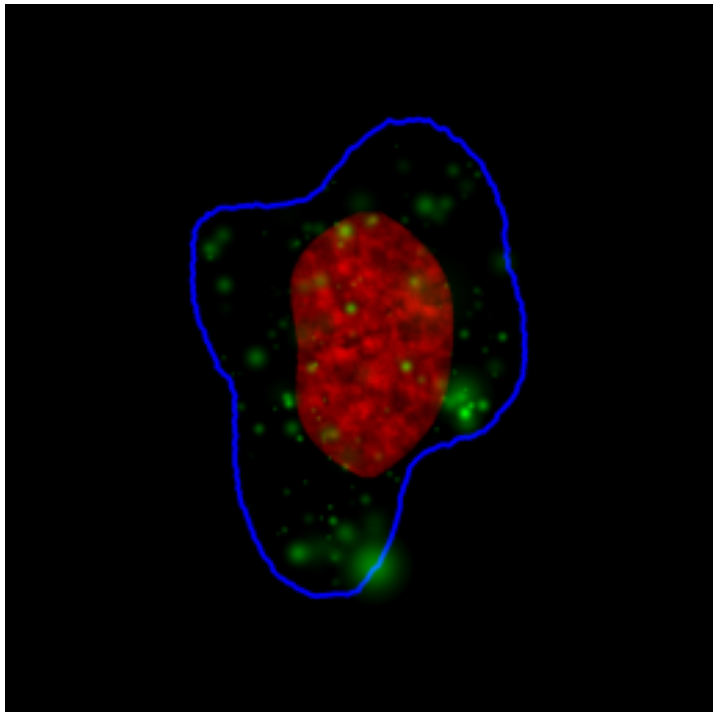


Zhao & Murphy 2007

Synthesized Images



Lysosomes



Endosomes

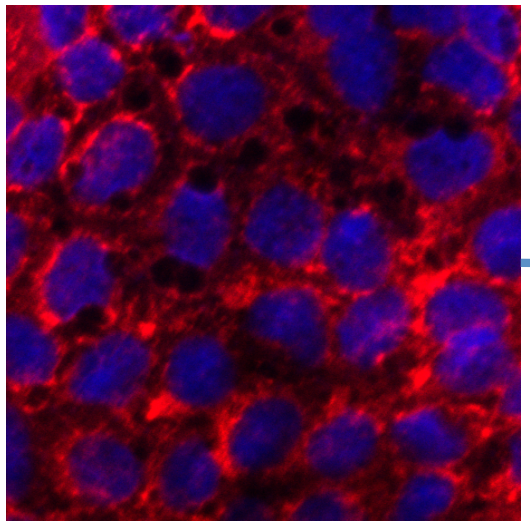
- SLML toolbox - Ivan Cao-Berg, Tao Peng, Ting Zhao
- Have portable tool for generating images from model

SLML Suite and OMERO

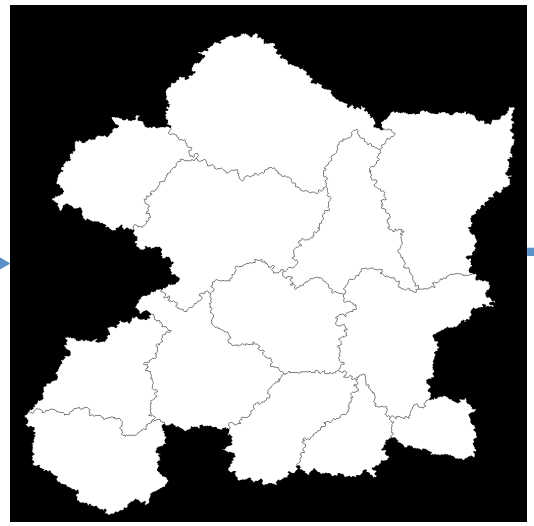
- Adding same OMERO read interface used by PUnMix

Meta Classification: Graphical Models

Original image



segmentation



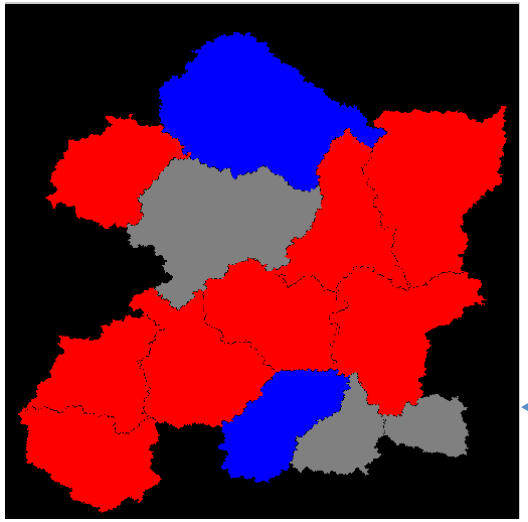
feature extraction

Cell#	feat1	feat2	feat3 …
1	0.78	15.7	0.01
2	0.83	23.6	0.01
3	0.71	20.9	0.03
…			

Trained classifier

likelihood of each class

Cell#	ER	Lyso.	Tub…
1	**0.58**	0.32	0.10
2	0.40	**0.45**	0.15
3	0.32	0.30	**0.38**
…			

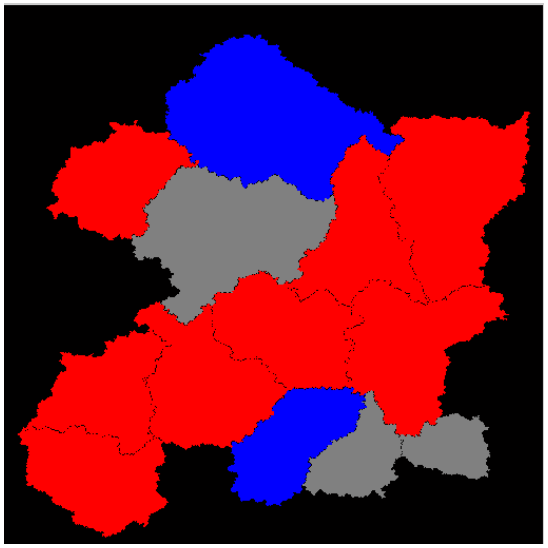


Result of the classification

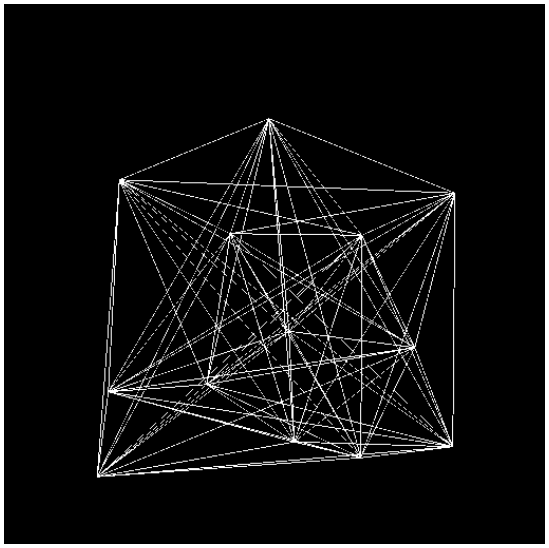
ER
Lysosome
Tubulin

Improve labels by loopy belief propagation

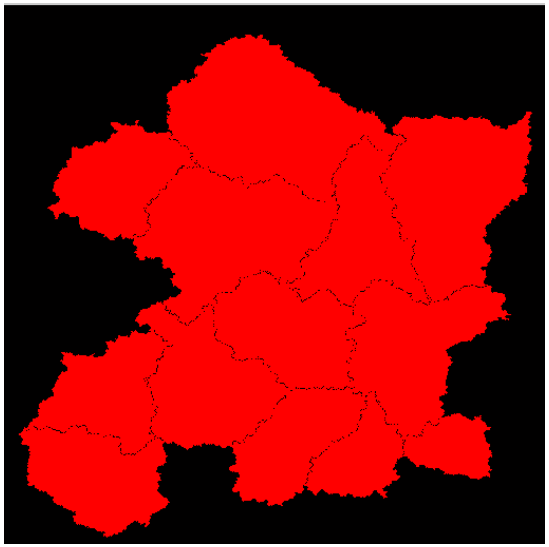
Confusion Matrix after prior updating



Result of the classification



Graph



Prior updating

	Nuclear	ER	Golgi	Lyso.	Mito.	Nucleol	Actin	Endo.	Tubulin
Nuclear	**100**	0	0	0	0	0	0	0	0
ER	0	**98.8**	0	0.8	0.4	0	0	0	0
Golgi	0	0.2	**93.9**	0.7	5.2	0	0	0	0
Lysosomal	0	0.2	0	**99.8**	0	0	0	0	0
Mito.	0	0	0	1.6	**98.4**	0	0	0	0
Nucleolar	0	0	0	0	0	**100**	0	0	0
Actin	0	3.0	0	0	0	0	**97.0**	0	0
Endosom.	0	5.6	0	12.2	3.8	0	5.6	**72.5**	0.3
Tubulin	0	0.3	0	0	0	0	0	0	**99.7**

- Overall accuracy to recognize the 9 classes = 95.6%

Acknowledgements

- Dr. Michael Boland
- Dr. Meel Velliste
- Dr. Kai Huang
- Dr. Xiang Chen
- Dr. Ting Zhao
- Dr. Shann-Ching Chen
- Dr. Juchang Hua
- Dr. Justin Newberg
- Dr. Baek Hwan Cho
- Tao Peng
- Jenn Backal

Funding
- NSF, NIH, Commonwealth of Pennsylvania

Collaborators/Consultants
- David Casasent, Simon Watkins, Jon Jarvik, Peter Berget, Jack Rohrer, Tom Mitchell, Christos Faloutsos, Jelena Kovacevic, William Cohen, Geoff Gordon