Requirements for Persistent Identifiers within the MPI-CBG

Jeff Oegema
Max Planck Institute of Molecular Cell Biology and Genetics

eScience Seminar, March 28th, 2008

The Question

PIDs are currently used:

- for publications (ex: CrossRef)
- within archives and libraries
- as a tool for primary scientific data (World Data Center for Climate)
- in a variety of other ways

What benefits can PIDs bring to the types of data that the MPI-CBG generates?

CBG

Max Planck Institute
of Molecular Cell Biology
and Genetics

Outline

1) MPI-CBG Data “Objects” and Examples

2) Current State of Data Handling and Software within the MPI-CBG

3) Benefits from PIDs and potential applications and questions

CBG
Max Planck Institute
of Molecular Cell Biology
and Genetics

Types of Data “Objects” Gathered

Microscopy Data - Fluorescent, Confocal, High-Throughput Screening, Electron Microscopy

Other Image Data - Protein Blots, Electrophoresis Gels, etc.

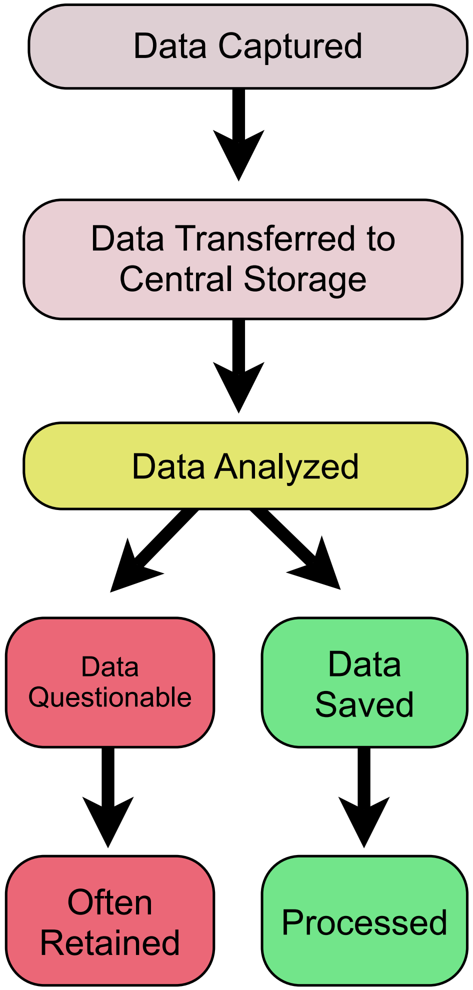
Sequence Data - DNA Sequencing Results

Mass Spec Data- Spectra and Numerical Data Files

Protocols and Methods

The MPI-CBG is heavily focused on microscopy and imaging which accounts for **95%** of our data by volume

MPI-CBG Data Example



In most / many cases data captured is an image or movie as a result from an experiment

Structure is managed by the scientist or research group - file server space allocated by project

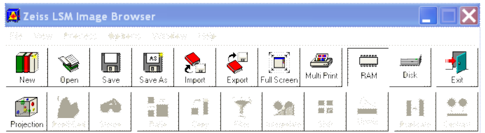
Data is analyzed and evaluated, often manually

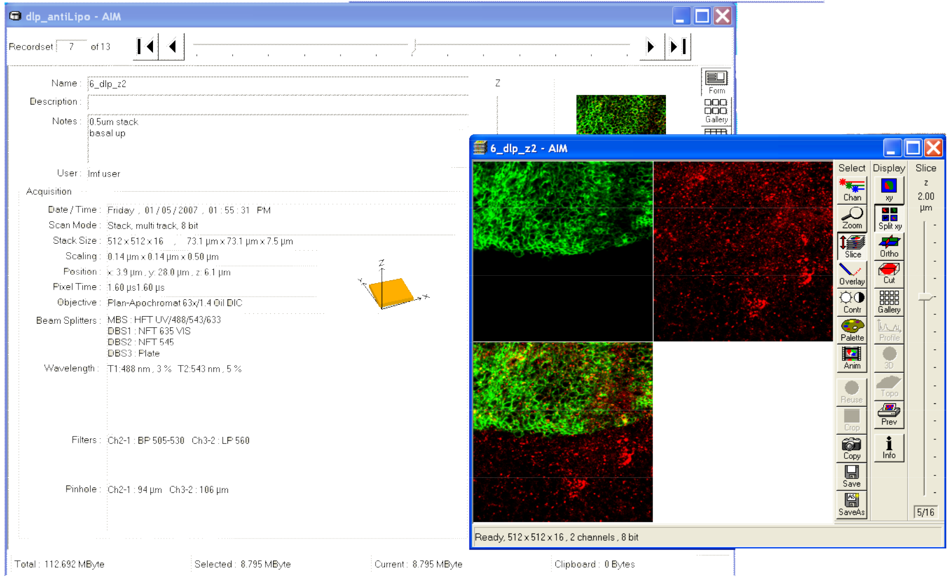
After analysis the data is deemed to be valuable, a failure, or the results are indeterminate.

It is often unclear for quite some time after capture if the data has meaning or value.

What does the data look like?

Primary Data





Recordset 7 of 13

Name : 6_dlp_z2
Description :
Notes : 0.5um stack
basal up
User : lmf user

Acquisition
Date / Time : Friday , 01 / 05 / 2007 , 01 : 55 : 31 PM
Scan Mode : Stack, multi track, 8 bit
Stack Size : 512 x 512 x 16 , 73.1 µm x 73.1 µm x 7.5 µm
Scaling : 0.14 µm x 0.14 µm x 0.50 µm
Position : x: 3.9 µm , y: 28.0 µm , z: 6.1 µm
Pixel Time : 1.60 µs1.60 µs
Objective : Plan-Apochromat 63x/1.4 Oil DIC
Beam Splitters : MBS : HFT UV/488/543/633
DBS1 : NFT 635 VIS
DBS2 : NFT 545
DBS3 : Plate
Wavelength : T1:488 nm , 3 % T2:543 nm , 5 %
Filters : Ch2-1 : BP 505-530 Ch3-2 : LP 560
Pinhole : Ch2-1 : 94 µm Ch3-2 : 106 µm

Total : 112.692 MByte Selected : 8.795 MByte Current : 8.795 MByte Clipboard : 0 Bytes

Ready, 512 x 512 x 16 , 2 channels , 8 bit

Christina Eugster - Eaton Lab

In Zeiss LSM file format (Zeiss Meta Imaging System)

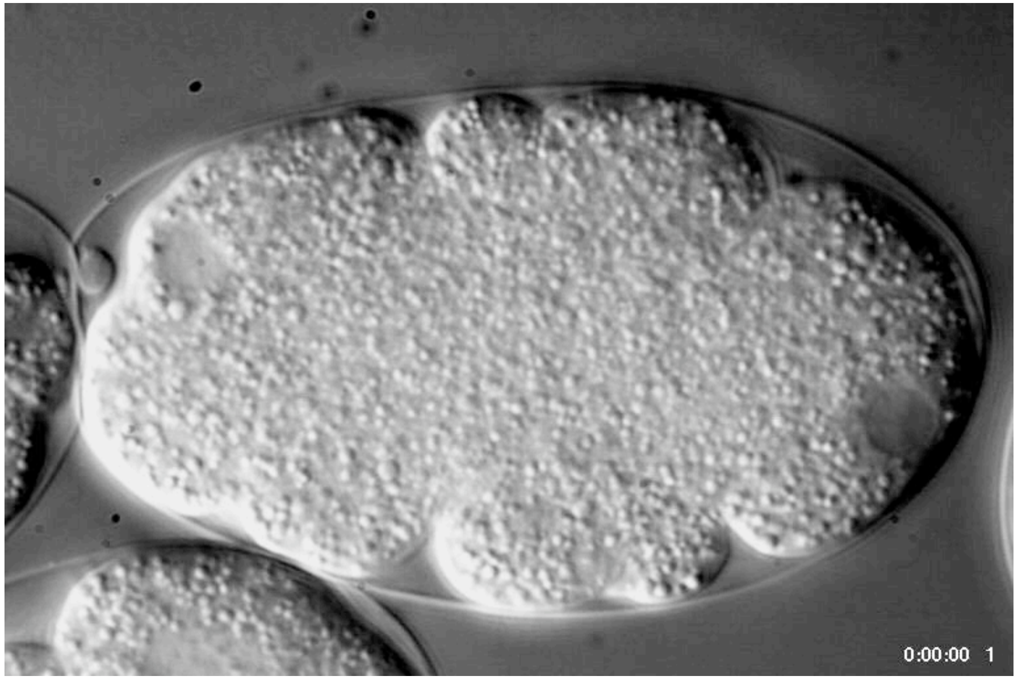
Readable using the Zeiss software or an ImageJ Plugin

This image is of larval tissue of a fruit fly and shows the expression of the Dally protein in green and lipoprotein in red.

Dally is overexpressed.

What does the data look like?

Primary Data



C. Elegans Division - Hyman Lab

Captured using normal contrast microscopy

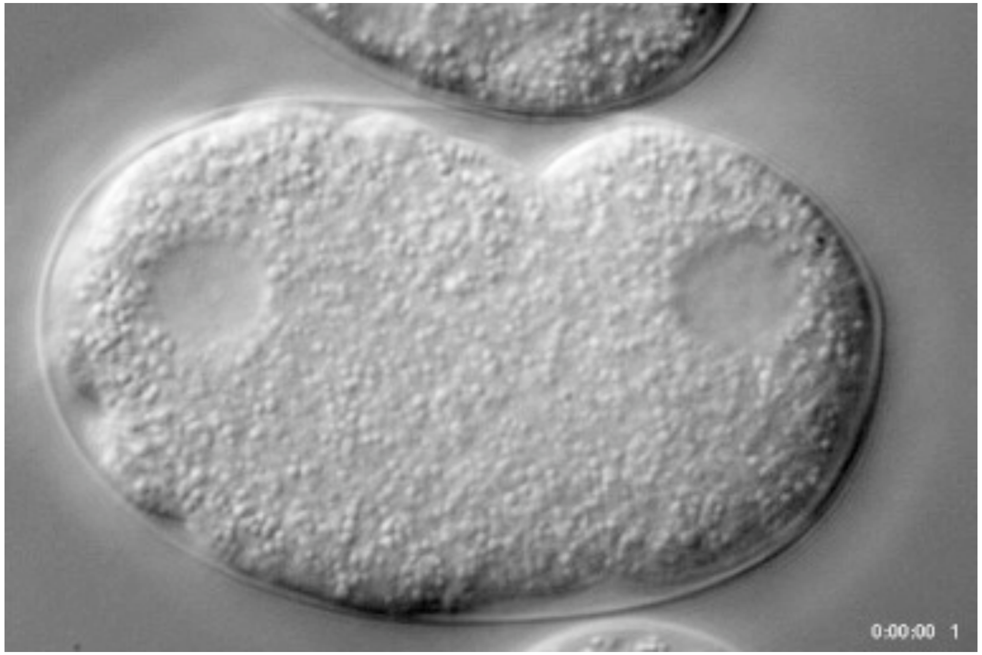
Uncompressed movie format

No metadata encoded with file on capture, but was manually added

The movie shows the merging of the egg and sperm cells and the first divisions of a C. Elegans embryo

What does the data look like?

Primary Data



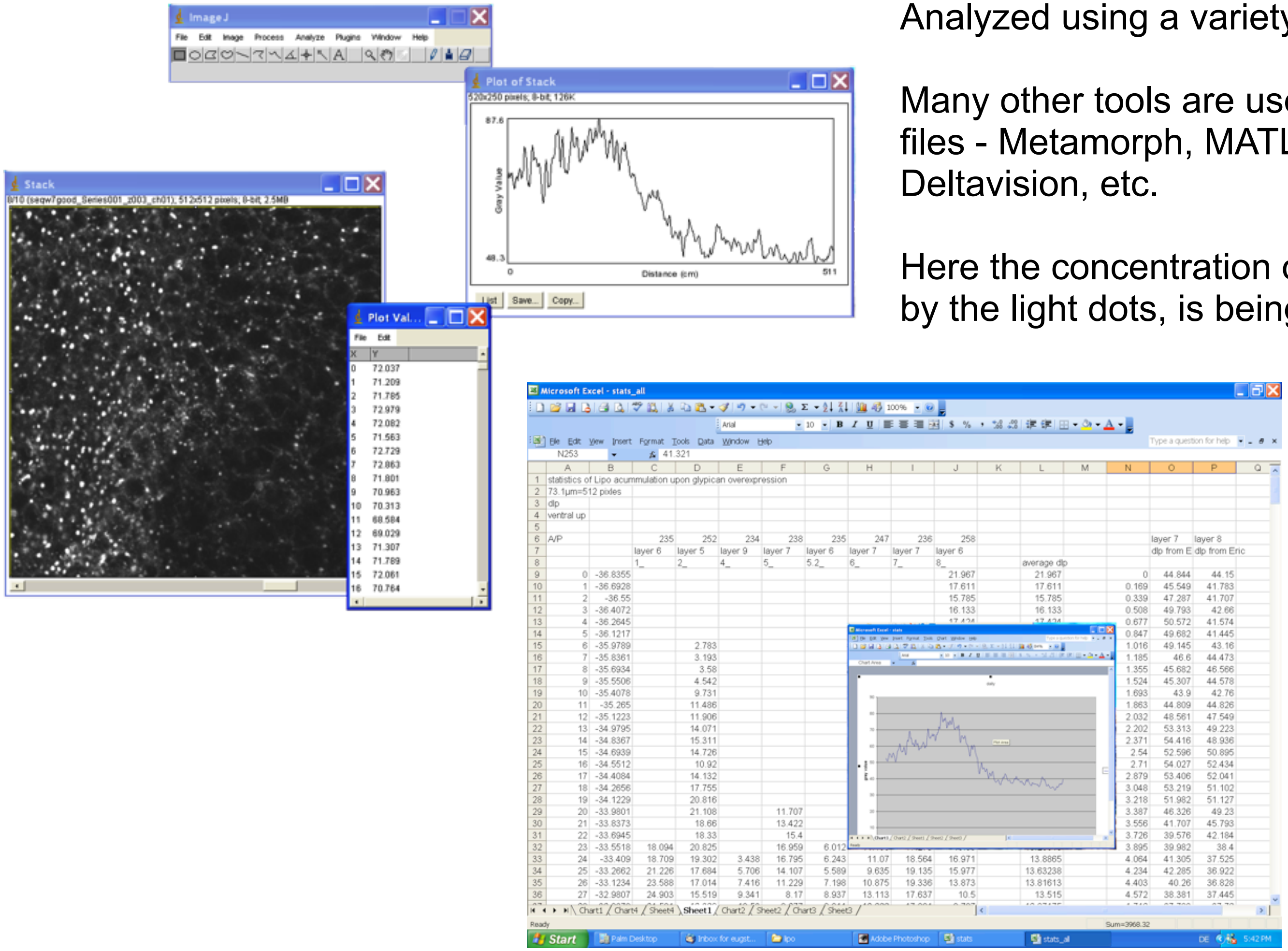
C. Elegans Division - Hyman Lab

This movie shows the failure pattern when gene H04J21.3 was knocked down.

This pattern was interpreted by eye and was classified as a "Spindle Assembly" problem

What does the data look like?

Analysis of Data


ImageJ
File Edit Image Process Analyze Plugins Window Help
Plot of Stack
520x250 pixels; 8-bit; 126K
87.6
48.3
Gray Value
0
Distance (cm)
511
List
Save...
Copy...
Stack
8/10 (seqw7good_Series001_z003_ch01); 512x512 pixels; 8-bit; 2.5MB
Plot Val...
File Edit
Microsoft Excel - stats_all
File Edit View Insert Format Tools Data Window Help
Type a question for help
Arial
10
100%
N253
41.321
Microsoft Excel - stats
Chart Area
Plot Area
Chart1 Chart4 Sheet4 Sheet1 Chart2 Sheet2 Chart3 Sheet3
Ready
Sum=3968.32
Start
Palm Desktop
Inbox for eugst...
lpo
Adobe Photoshop
stats
stats_all
DE
5:42 PM

X	Y
0	72.037
1	71.209
2	71.785
3	72.979
4	72.082
5	71.563
6	72.729
7	72.863
8	71.801
9	70.963
10	70.313
11	68.584
12	69.029
13	71.307
14	71.789
15	72.061
16	70.764

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	statistics of Lipo acummulation upon glypican overexpression																
2	73.1µm=512 pixles																
3	dlp																
4	ventral up																
5																	
6	A/P		235	252	234	238	235	247	236	258					layer 7	layer 8	
7			layer 6	layer 5	layer 9	layer 7	layer 6	layer 7	layer 7	layer 6					dlp from E	dlp from Eric	
8			1_	2_	4_	5_	5.2_	6_	7_	8_		average dlp					
9	0	-36.8355								21.967		21.967		0	44.844	44.15	
10	1	-36.6928								17.611		17.611		0.169	45.549	41.783	
11	2	-36.55								15.785		15.785		0.339	47.287	41.707	
12	3	-36.4072								16.133		16.133		0.508	49.793	42.66	
13	4	-36.2645								17.424		17.424		0.677	50.572	41.574	
14	5	-36.1217												0.847	49.682	41.445	
15	6	-35.9789		2.783										1.016	49.145	43.16	
16	7	-35.8361		3.193										1.185	46.6	44.473	
17	8	-35.6934		3.58										1.355	45.682	46.566	
18	9	-35.5506		4.542										1.524	45.307	44.578	
19	10	-35.4078		9.731										1.693	43.9	42.76	
20	11	-35.265		11.486										1.863	44.809	44.826	
21	12	-35.1223		11.906										2.032	48.561	47.549	
22	13	-34.9795		14.071										2.202	53.313	49.223	
23	14	-34.8367		15.311										2.371	54.416	48.936	
24	15	-34.6939		14.726										2.54	52.596	50.895	
25	16	-34.5512		10.92										2.71	54.027	52.434	
26	17	-34.4084		14.132										2.879	53.406	52.041	
27	18	-34.2656		17.755										3.048	53.219	51.102	
28	19	-34.1229		20.816										3.218	51.982	51.127	
29	20	-33.9801		21.108		11.707								3.387	46.326	49.23	
30	21	-33.8373		18.66		13.422								3.556	41.707	45.793	
31	22	-33.6945		18.33		15.4								3.726	39.576	42.184	
32	23	-33.5518	18.094	20.825		16.959	6.012							3.895	39.982	38.4	
33	24	-33.409	18.709	19.302	3.438	16.795	6.243	11.07	18.564	16.971		13.8865		4.064	41.305	37.525	
34	25	-33.2662	21.226	17.684	5.706	14.107	5.589	9.635	19.135	15.977		13.63238		4.234	42.285	36.922	
35	26	-33.1234	23.588	17.014	7.416	11.229	7.198	10.875	19.336	13.873		13.81613		4.403	40.26	36.828	
36	27	-32.9807	24.903	15.519	9.341	8.17	8.937	13.113	17.637	10.5		13.515		4.572	38.381	37.445	

Analyzed using a variety of tools - Example ImageJ

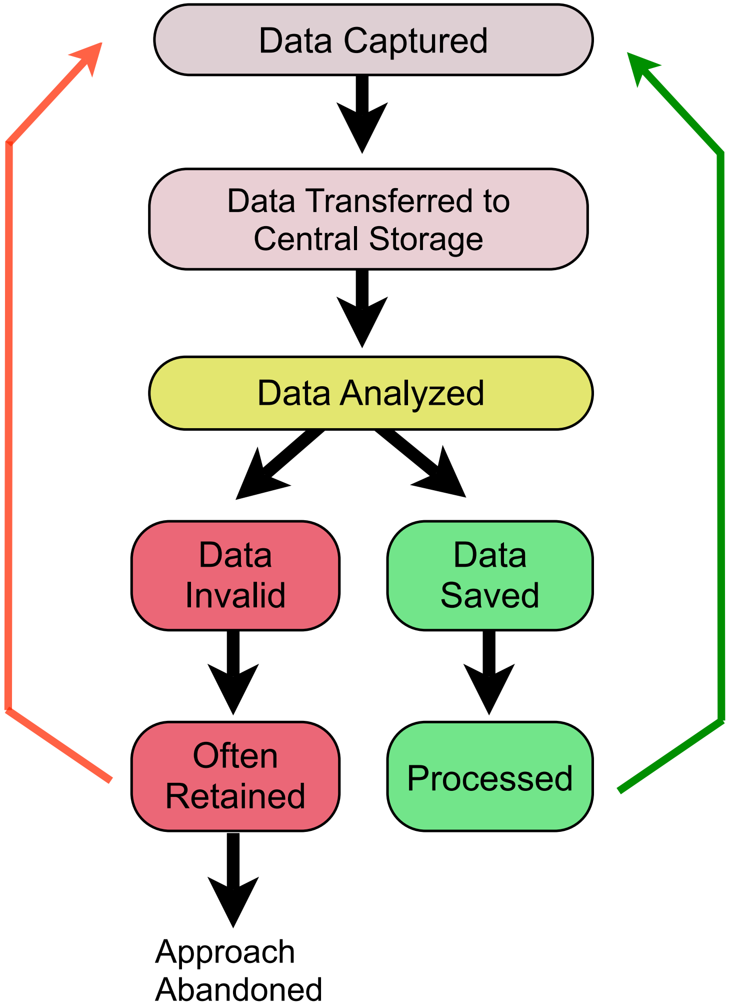
Many other tools are used and generate a variety of files - Metamorph, MATLAB, Huygens, Definiens, Deltavision, etc.

Here the concentration of lipoprotein, as indicated by the light dots, is being examined and plotted.

What does the data look like?

Experimental Repeats

Change of:
Experimental Conditions
Protocol
Capture Method



Data Captured

Data Transferred to Central Storage

Data Analyzed

Data Invalid

Data Saved

Often Retained

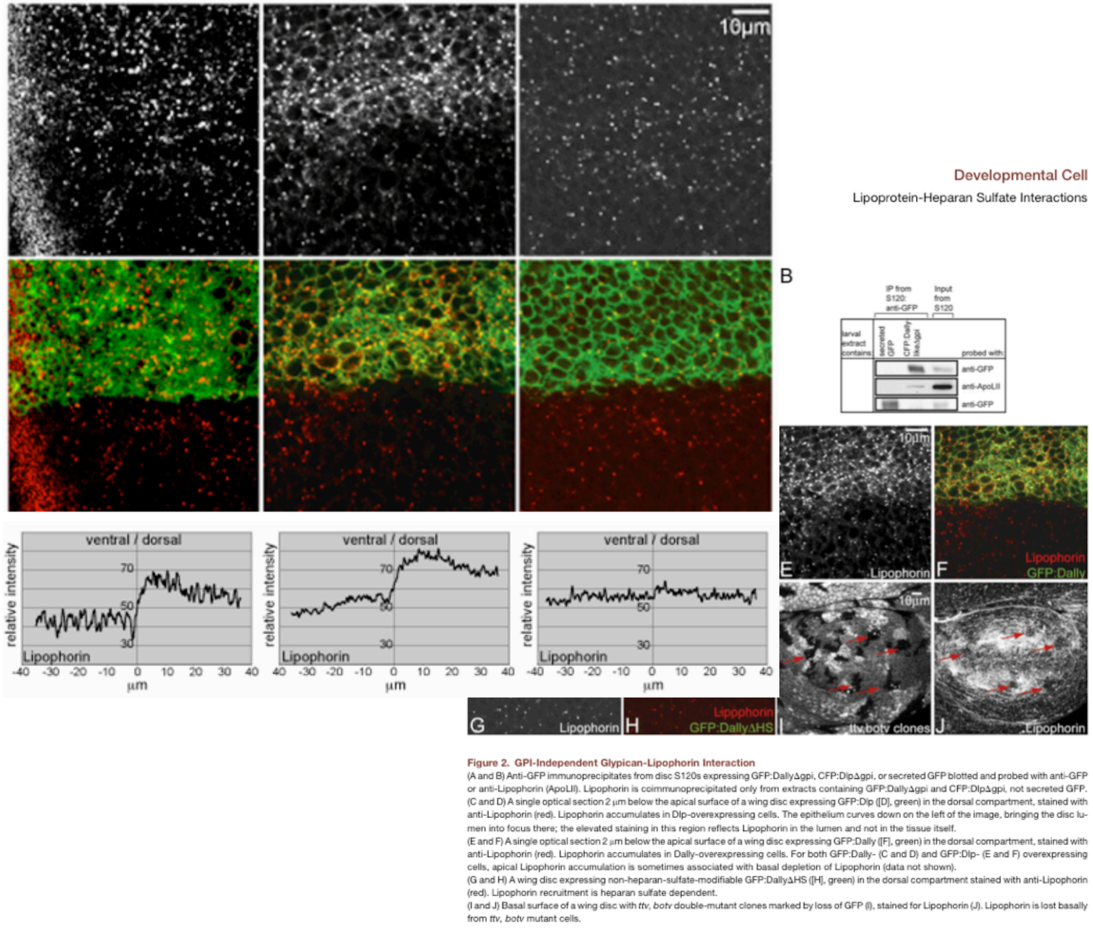
Processed

Approach Abandoned

Repetition for:
Confirmation / Statistics
Image Quality

What does the data look like?

Figures and Publication



Developmental Cell

Lipoprotein-Heparan Sulfate Interactions

Figure 2. GPI-Independent Glypican-Lipophorin Interaction

(A and B) Anti-GFP immunoprecipitates from disc S120s expressing GFP:Dally∆gpi, CFP:Dlp∆gpi, or secreted GFP blotted and probed with anti-GFP or anti-Lipophorin (ApoLII). Lipophorin is coimmunoprecipitated only from extracts containing GFP:Dally∆gpi and CFP:Dlp∆gpi, not secreted GFP.

(C and D) A single optical section 2 µm below the apical surface of a wing disc expressing GFP:Dlp ([D], green) in the dorsal compartment, stained with anti-Lipophorin (red). Lipophorin accumulates in Dlp-overexpressing cells. The epithelium curves down on the left of the image, bringing the disc lumen into focus there; the elevated staining in this region reflects Lipophorin in the lumen and not in the tissue itself.

(E and F) A single optical section 2 µm below the apical surface of a wing disc expressing GFP:Dally ([F], green) in the dorsal compartment, stained with anti-Lipophorin (red). Lipophorin accumulates in Dally-overexpressing cells. For both GFP:Dally- (C and D) and GFP:Dlp- (E and F) overexpressing cells, apical Lipophorin accumulation is sometimes associated with basal depletion of Lipophorin (data not shown).

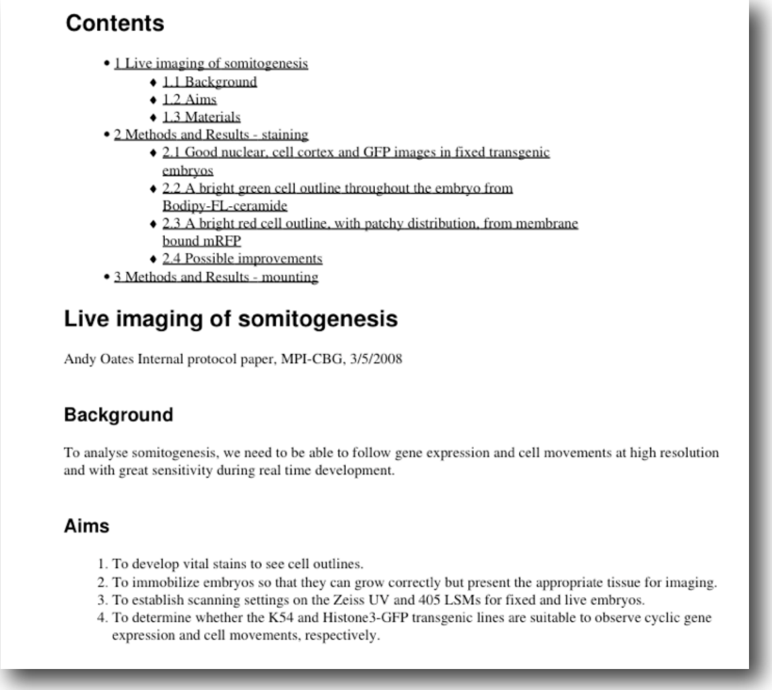
(G and H) A wing disc expressing non-heparan-sulfate-modifiable GFP:Dally∆HS ([H], green) in the dorsal compartment stained with anti-Lipophorin (red). Lipophorin recruitment is heparan sulfate dependent.

(I and J) Basal surface of a wing disc with *ttv*, *botv* double-mutant clones marked by loss of GFP (I), stained for Lipophorin (J). Lipophorin is lost basally from *ttv*, *botv* mutant cells.

Multiple experiments and primary data sets come together for a single publication figure.

Here the concentrations of several proteins are being shown side by side.

Complications



Contents

Live imaging of somitogenesis

Andy Oates Internal protocol paper, MPI-CBG, 3/5/2008

Background

To analyse somitogenesis, we need to be able to follow gene expression and cell movements at high resolution and with great sensitivity during real time development.

Aims

1. To develop vital stains to see cell outlines.
2. To immobilize embryos so that they can grow correctly but present the appropriate tissue for imaging.
3. To establish scanning settings on the Zeiss UV and 405 LSMs for fixed and live embryos.
4. To determine whether the K54 and Histone3-GFP transgenic lines are suitable to observe cyclic gene expression and cell movements, respectively.

Protocols change, sometimes with every repetition

A proper (and complete) description must be bound to each data set or the data is uninterpretable

Screening Data

Automation of Data Generation:

- Simplifies protocol data as one protocol is used for each sample
- Produces millions of image and analysis results

Image courtesy of Tecan Group Ltd.

- Data collection is automated
- Analysis may be automated

Screening Example - Genome Wide Screen

Data Capture	19.1 TB of Image Data (tiff), over 2,000,000 images
Analysis	Thousands of computer hours were used to generate numerical analysis
Interpretation	Much of the data still human interpreted based on the numerical analysis

Analysis is only as good as the image processing tools used

Human interpretation is based on current knowledge and limited by human perception.

Data Summary

Experimental Data

Generated by a specific protocol

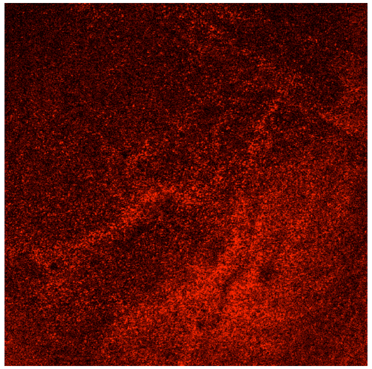
Meant to answer a single question

May not have additional information

Generally is image or movie data

Analysis generates numerical data

Almost always repeated, sometimes with protocol changes



Lipoprotein Distribution - Eaton Lab

Screening Data

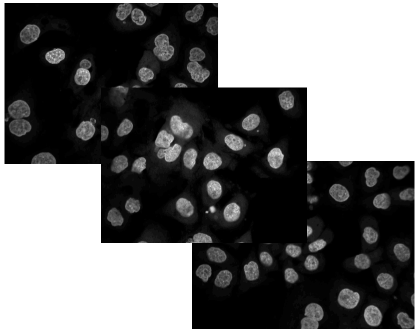
Before going to screening the protocol must be locked

Experimentation is automated

Analysis is computationally intensive

Still usually requires human interpretation

Is likely to be useful for future reanalysis



GWS Imaging - Zerial Lab

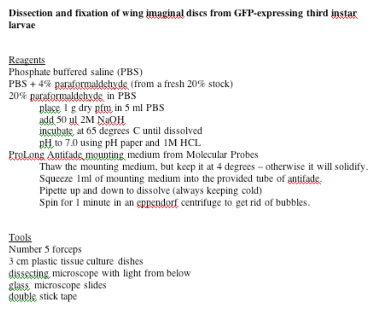
Protocols and Methods

A protocol is linked to an experiment, or to a screen

The format of the protocols kept varies highly

Experimental data is only meaningful in combination with the protocol

Protocols themselves have value as they teach methods to scientists



Dissection and fixation of wing imaginal discs from GFP-expressing third instar larvae

Reagents
Phosphate buffered saline (PBS)
PBS + 4% paraformaldehyde (from a fresh 20% stock)
20% paraformaldehyde in PBS
    place 1 g dry pfm in 5 ml PBS
    add 50 ul 2M NaOH
    incubate at 65 degrees C until dissolved
    pH to 7.0 using pH paper and 1M HCL
ProLong Antifade mounting medium from Molecular Probes
    Thaw the mounting medium, but keep it at 4 degrees – otherwise it will solidify.
    Squeeze 1ml of mounting medium into the provided tube of antifade.
    Pipette up and down to dissolve (always keeping cold)
    Spin for 1 minute in an eppendorf centrifuge to get rid of bubbles.

Tools
Number 5 forceps
3 cm plastic tissue culture dishes
dissecting microscope with light from below
glass microscope slides
double stick tape

Staining Protocol - Eaton Lab

Data Summary

Data within the MPI-CBG is captured and organized in many ways.

Before we can present primary data to the world we must improve our own internal data handling.

Data within the building must be consistent, uniquely identified, and associated with the metadata.

If PIDs are ready for us, are we ready for them?

Current Efforts

Improvement of internal standards

Professionalization of software development for data management (including internal metadata and identifiers)

Data management migrating out of individual non-standard efforts towards standards-compliant systems provided by a central software development group

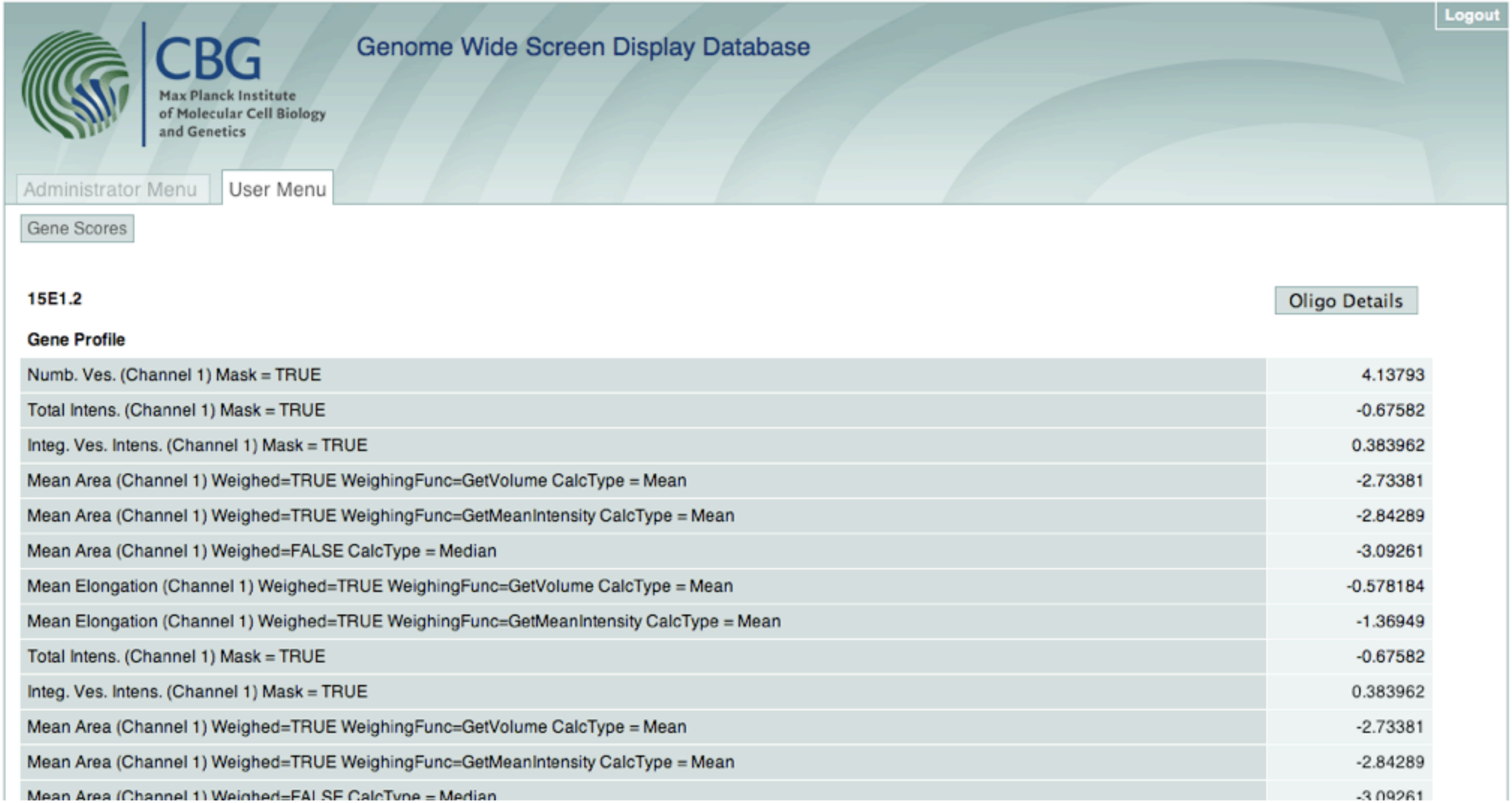
Formation of an image processing and analysis facility to provide central know-how and tools

Current Efforts

Specific “display” database systems provide access to information for a given screen or publication when this is considered necessary

Data management systems are being created to assist with information flow and experiment management

Protocols are currently managed by wiki, in a simple database system, or manually, depending on group. Plans exist to standardize format, structure, and location of these protocols.

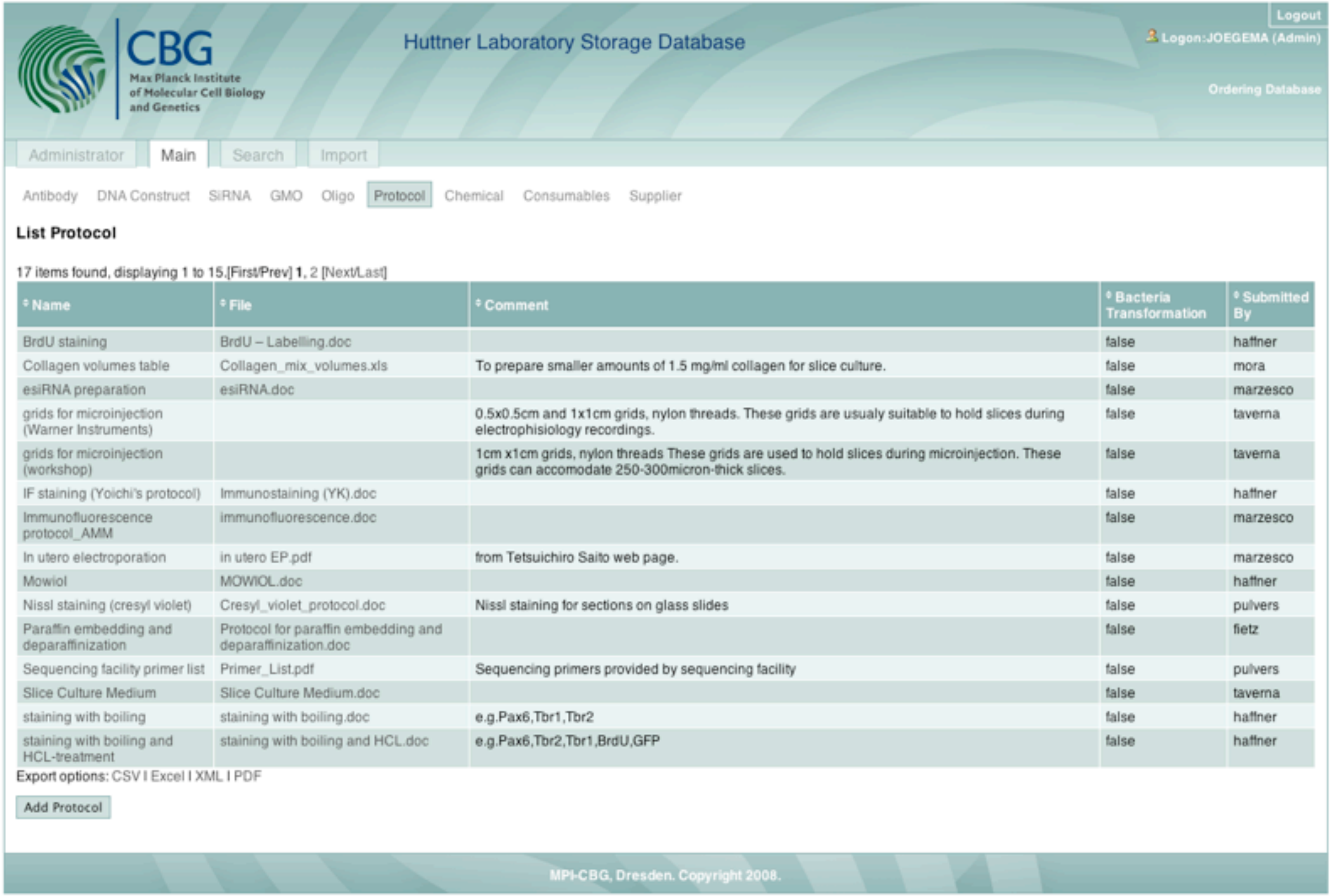


Genome Wide Screen Display Database

15E1.2

Gene Profile

Numb. Ves. (Channel 1) Mask = TRUE	4.13793
Total Intens. (Channel 1) Mask = TRUE	-0.67582
Integ. Ves. Intens. (Channel 1) Mask = TRUE	0.383962
Mean Area (Channel 1) Weighed=TRUE WeighingFunc=GetVolume CalcType = Mean	-2.73381
Mean Area (Channel 1) Weighed=TRUE WeighingFunc=GetMeanIntensity CalcType = Mean	-2.84289
Mean Area (Channel 1) Weighed=FALSE CalcType = Median	-3.09261
Mean Elongation (Channel 1) Weighed=TRUE WeighingFunc=GetVolume CalcType = Mean	-0.578184
Mean Elongation (Channel 1) Weighed=TRUE WeighingFunc=GetMeanIntensity CalcType = Mean	-1.36949
Total Intens. (Channel 1) Mask = TRUE	-0.67582
Integ. Ves. Intens. (Channel 1) Mask = TRUE	0.383962
Mean Area (Channel 1) Weighed=TRUE WeighingFunc=GetVolume CalcType = Mean	-2.73381
Mean Area (Channel 1) Weighed=TRUE WeighingFunc=GetMeanIntensity CalcType = Mean	-2.84289
Mean Area (Channel 1) Weighed=FALSE CalcType = Median	-3.09261



Huttner Laboratory Storage Database

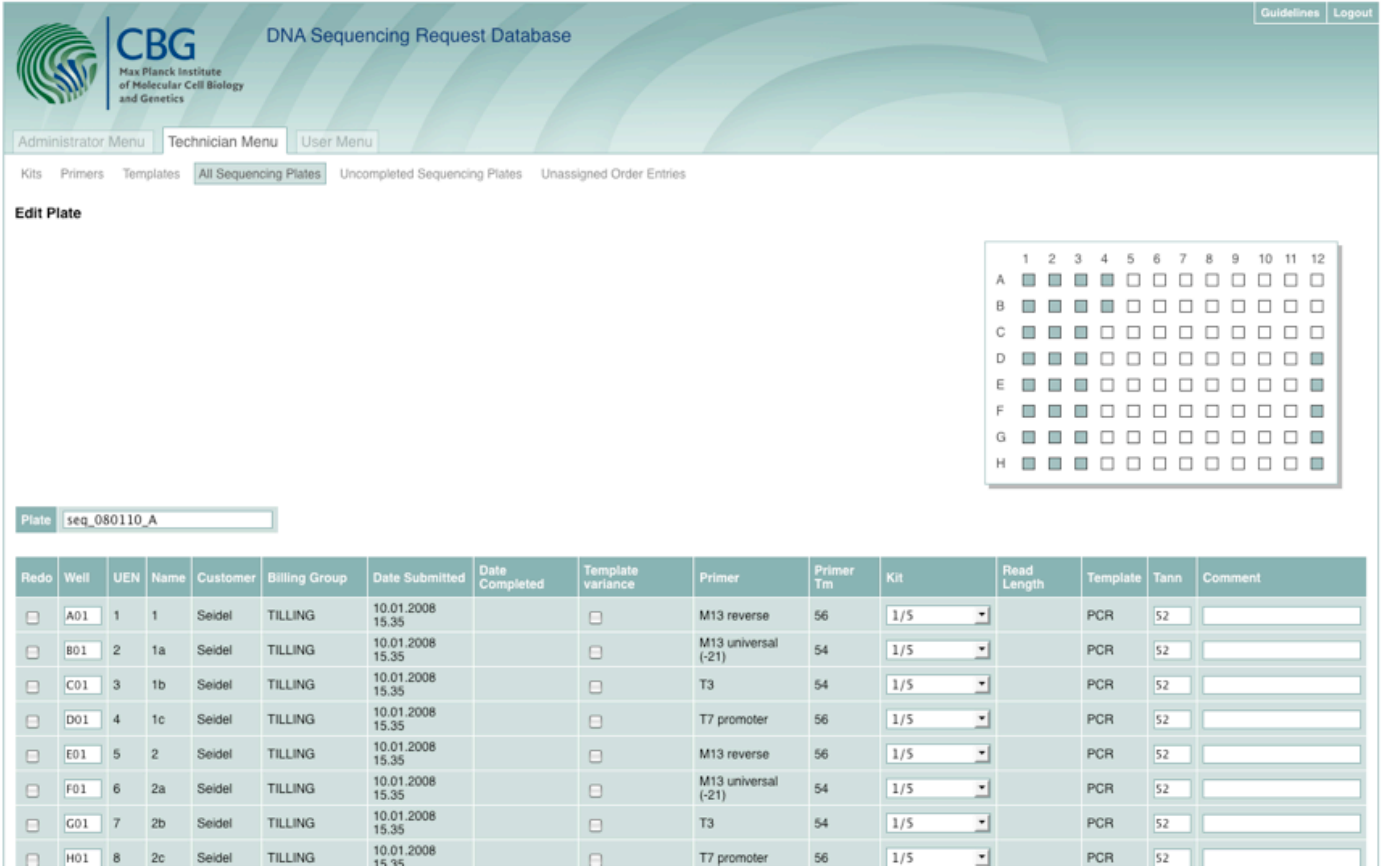
List Protocol

17 items found, displaying 1 to 15.[First/Prev] 1, 2 [Next/Last]

Name	File	Comment	Bacteria Transformation	Submitted By
BrdU staining	BrdU – Labelling.doc		false	haffner
Collagen volumes table	Collagen_mix_volumes.xls	To prepare smaller amounts of 1.5 mg/ml collagen for slice culture.	false	mora
esiRNA preparation	esiRNA.doc		false	marzesco
grids for microinjection (Warner Instruments)		0.5x0.5cm and 1x1cm grids, nylon threads. These grids are usualy suitable to hold slices during electrophisiology recordings.	false	taverna
grids for microinjection (workshop)		1cm x1cm grids, nylon threads These grids are used to hold slices during microinjection. These grids can accomodate 250-300micron-thick slices.	false	taverna
IF staining (Yoichi's protocol)	Immunostaining (YK).doc		false	haffner
Immunofluorescence protocol_AMM	immunofluorescence.doc		false	marzesco
In utero electroporation	in utero EP.pdf	from Tetsuichiro Saito web page.	false	marzesco
Mowiol	MOWIOL.doc		false	haffner
Nissl staining (cresyl violet)	Cresyl_violet_protocol.doc	Nissl staining for sections on glass slides	false	pulvers
Paraffin embedding and deparaffinization	Protocol for paraffin embedding and deparaffinization.doc		false	fietz
Sequencing facility primer list	Primer_List.pdf	Sequencing primers provided by sequencing facility	false	pulvers
Slice Culture Medium	Slice Culture Medium.doc		false	taverna
staining with boiling	staining with boiling.doc	e.g.Pax6,Tbr1,Tbr2	false	haffner
staining with boiling and HCL-treatment	staining with boiling and HCL.doc	e.g.Pax6,Tbr2,Tbr1,BrdU,GFP	false	haffner

Export options: CSV I Excel I XML I PDF

MPI-CBG, Dresden. Copyright 2008.



DNA Sequencing Request Database

Edit Plate

	1	2	3	4	5	6	7	8	9	10	11	12
A	■	■	■	■	□	□	□	□	□	□	□	□
B	■	■	■	■	□	□	□	□	□	□	□	□
C	■	■	■	□	□	□	□	□	□	□	□	□
D	■	■	■	□	□	□	□	□	□	□	□	■
E	■	■	■	□	□	□	□	□	□	□	□	■
F	■	■	■	□	□	□	□	□	□	□	□	■
G	■	■	■	□	□	□	□	□	□	□	□	■
H	■	■	■	□	□	□	□	□	□	□	□	■

Plate: seq_080110_A

Redo	Well	UEN	Name	Customer	Billing Group	Date Submitted	Date Completed	Template variance	Primer	Primer Tm	Kit	Read Length	Template	Tann	Comment
□	A01	1	1	Seidel	TILLING	10.01.2008 15.35		□	M13 reverse	56	1/5		PCR	52	
□	B01	2	1a	Seidel	TILLING	10.01.2008 15.35		□	M13 universal (-21)	54	1/5		PCR	52	
□	C01	3	1b	Seidel	TILLING	10.01.2008 15.35		□	T3	54	1/5		PCR	52	
□	D01	4	1c	Seidel	TILLING	10.01.2008 15.35		□	T7 promoter	56	1/5		PCR	52	
□	E01	5	2	Seidel	TILLING	10.01.2008 15.35		□	M13 reverse	56	1/5		PCR	52	
□	F01	6	2a	Seidel	TILLING	10.01.2008 15.35		□	M13 universal (-21)	54	1/5		PCR	52	
□	G01	7	2b	Seidel	TILLING	10.01.2008 15.35		□	T3	54	1/5		PCR	52	
□	H01	8	2c	Seidel	TILLING	10.01.2008 15.35		□	T7 promoter	56	1/5		PCR	52	

PID Questions

The questions then become-

What role can and should PIDs play in our environment as we move towards internal standardization?

What role should they play in our distribution of data to a wider audience?

CBG

Max Planck Institute of Molecular Cell Biology and Genetics

PID Questions

- **Granularity** - At what level should a dataset be given a PID?

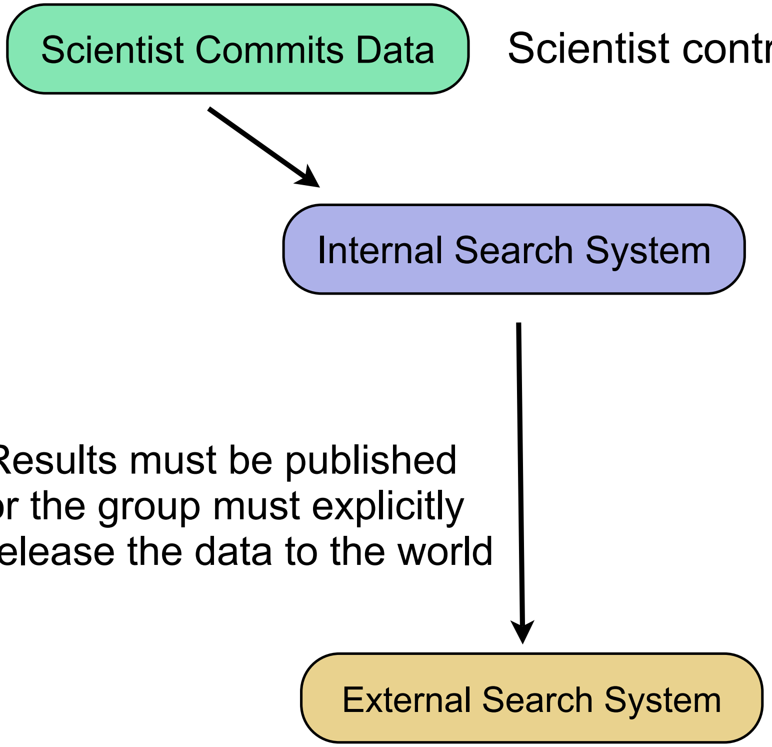
- **Timing** - When in the data life cycle should a dataset be given a PID? What do you do with multiple versions?

- **Access** - Should data at preliminary stages be accessible to the world or even to other internal groups? What should be done to address data privacy concerns, and data privacy requirements that vary over time?

- **Data Lockdown** - After a PID is assigned, how much can the data be changed?

- **Metadata** - What information should be associated with our PIDs, and are these schemas already existing or do we have to invent them?

Internal Search Tool



Scientist Commits Data

Scientist controls timing of original release

Internal Search System

Initially released internally
Data is tagged with metadata

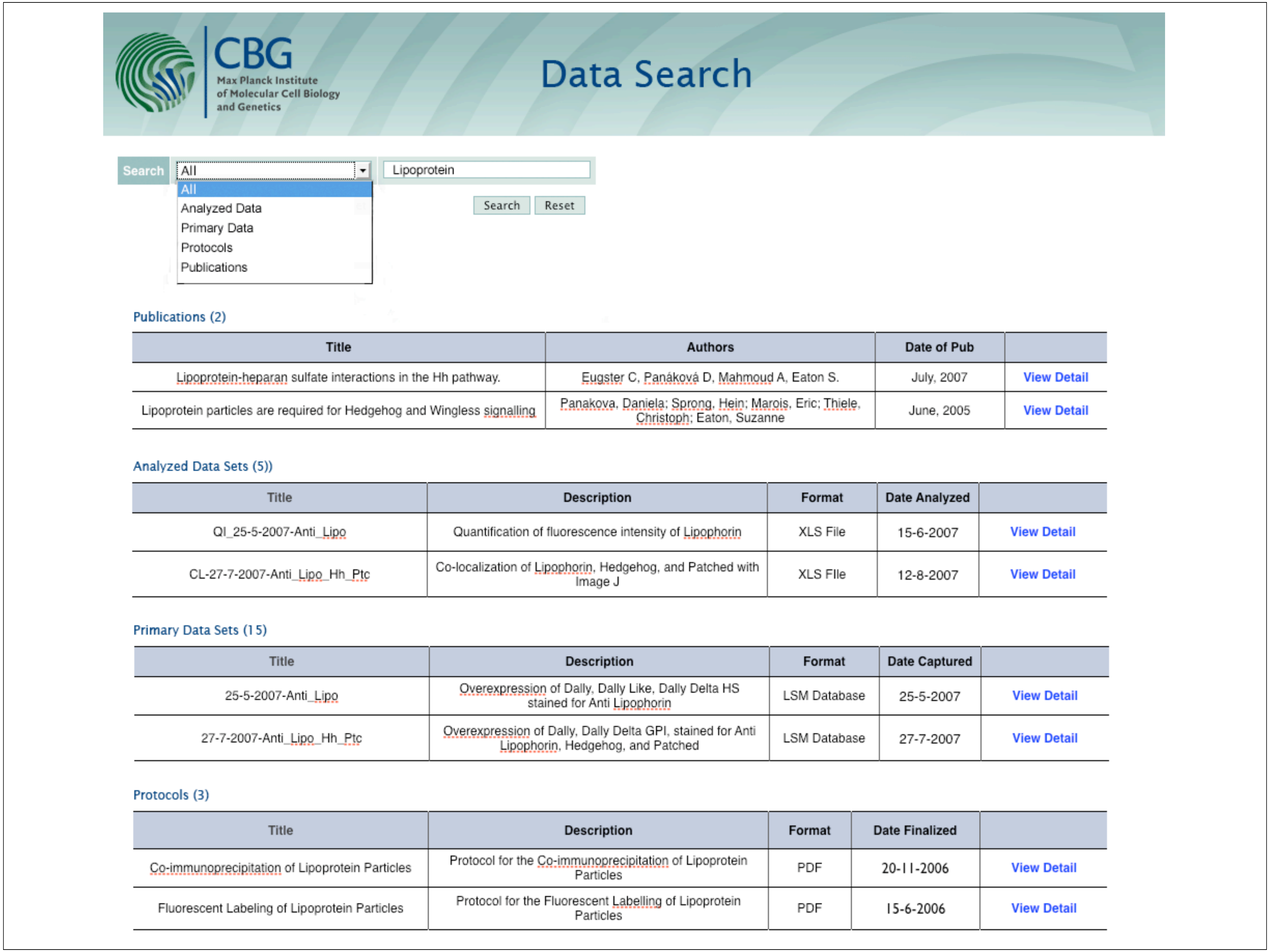
Results must be published
or the group must explicitly
release the data to the world

External Search System

Data goes to external search
and publication system

Data is tagged with a PID (DOI?)

Internal metadata should already
conform to standard



Data Search

Publications (2)

Title	Authors	Date of Pub	
Lipoprotein-heparan sulfate interactions in the Hh pathway.	Eugster C, Panáková D, Mahmoud A, Eaton S.	July, 2007	View Detail
Lipoprotein particles are required for Hedgehog and Wingless signalling	Panakova, Daniela; Sprong, Hein; Marois, Eric; Thiele, Christoph; Eaton, Suzanne	June, 2005	View Detail

Analyzed Data Sets (5))

Title	Description	Format	Date Analyzed	
QI_25-5-2007-Anti_Lipo	Quantification of fluorescence intensity of Lipophorin	XLS File	15-6-2007	View Detail
CL-27-7-2007-Anti_Lipo_Hh_Ptc	Co-localization of Lipophorin, Hedgehog, and Patched with Image J	XLS FIle	12-8-2007	View Detail

Primary Data Sets (15)

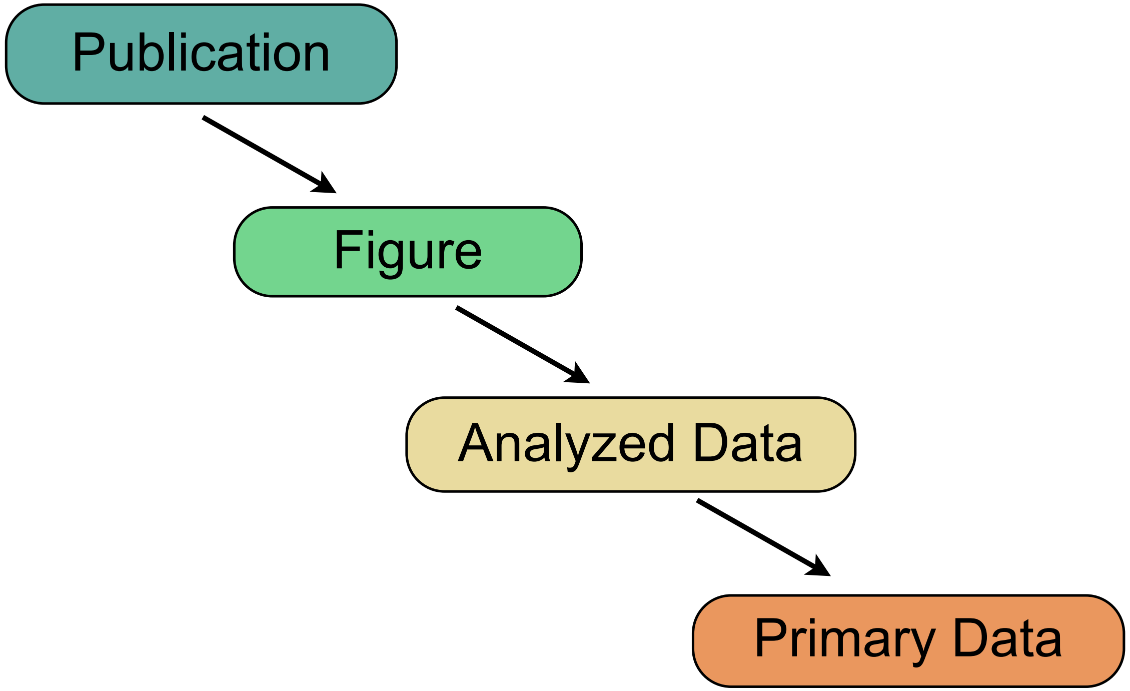
Title	Description	Format	Date Captured	
25-5-2007-Anti_Lipo	Overexpression of Dally, Dally Like, Dally Delta HS stained for Anti Lipophorin	LSM Database	25-5-2007	View Detail
27-7-2007-Anti_Lipo_Hh_Ptc	Overexpression of Dally, Dally Delta GPI, stained for Anti Lipophorin, Hedgehog, and Patched	LSM Database	27-7-2007	View Detail

Protocols (3)

Title	Description	Format	Date Finalized	
Co-immunoprecipitation of Lipoprotein Particles	Protocol for the Co-immunoprecipitation of Lipoprotein Particles	PDF	20-11-2006	View Detail
Fluorescent Labeling of Lipoprotein Particles	Protocol for the Fluorescent Labelling of Lipoprotein Particles	PDF	15-6-2006	View Detail

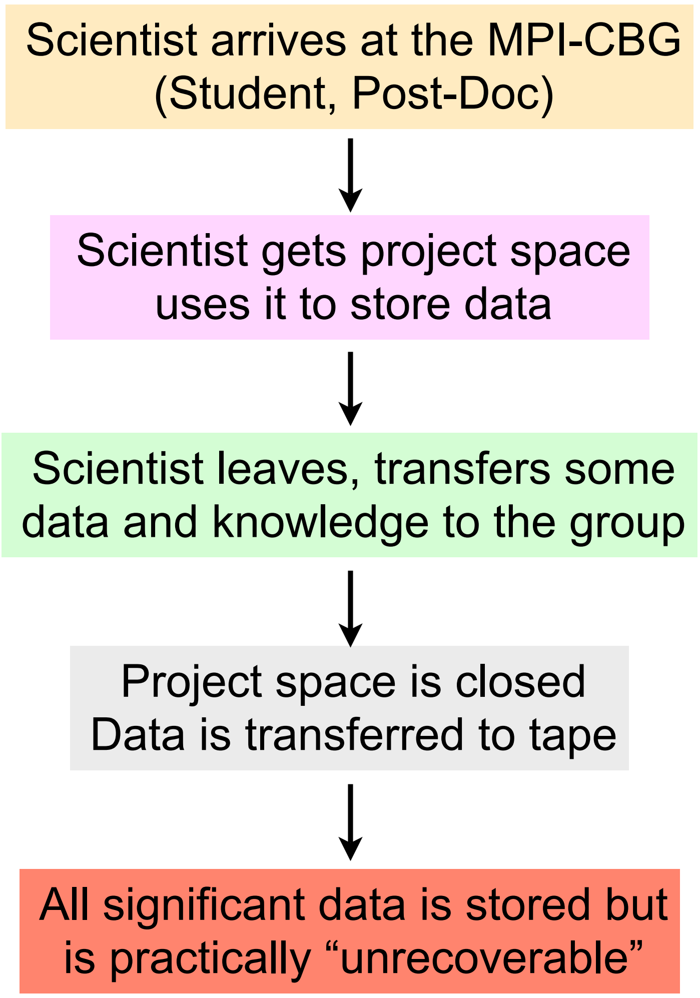
World Access

Publication Drill-Down



A scientist reading a publication has access to the primary data and can verify the correctness of the publication

Archiving and Data Loss



Scientist arrives at the MPI-CBG
(Student, Post-Doc)

Scientist gets project space
uses it to store data

Scientist leaves, transfers some
data and knowledge to the group

Project space is closed
Data is transferred to tape

All significant data is stored but
is practically “unrecoverable”

Conclusion

Benefits

Screening Data has the potential for future data mining to yield significant gain

Linking analyzed and primary data to publications allows greater transparency and validation of scientific work

Successful metadata binding to experimental data prevents data loss, allows greater reuse, and easier automated data mining

PIDs allow easier reuse of data at a later point, greater compatibility/interoperability with other systems and potentially other organizations with related data.

Difficulties

It is difficult to determine at what granularity and when to assign PIDs. This for us will require significant work on data organization structure.

We have only light experience with PIDs as a concept in-house at the moment.

Any system requiring the addition of metadata to primary research results must be extraordinarily easy to use or the research scientist won’t use it

Resources and funding within the MPI-CBG for such projects is difficult to obtain as the software facility and development resources are stretched with existing load.

Conclusion

The MPI-CBG is currently solidifying its internal data structures.

Until these structures are solid, its difficult to expand to greater world release of digital data through PIDs

The concept of PIDs and the associated metadata seems to have potential benefits for they types of data the MPI-CBG handles

The MPI-CBG will revisit the concept of PIDs as it’s software development progresses and appreciate feedback from the community.

Its likely there are tools and resources available that we are unaware of

Thanks!

Software Engineering Facility
Tim Cross - cross@mpi-cbg.de

Image Processing and Analysis Facility
Jean-Yves Tinevez - tinevez@mpi-cbg.de

IT Coordination
Jeffrey Oegema - joegema@mpi-cbg.de