PROCESSING...
Processing at SEMC

- Community
- Infrastructure
- Efficiency
- Balance
- Strategy
User profile

Breakdown of users
- Grad Students: 45%
- Post Docs: 21%
- PIs: 19%
- Professional Staff: 15%

Breakdown of projects
- Single particle: 70%
- 2dx/helical: 7%
- FIB-SEM: 19%
- Tomography: 3%
- Other: 1%

New York Structural Biology Center
Top requests for processing

Breakdown of projects

What is asked

- RELION / FREALIGN / cryoSPARC
- IMOD / Protomo / Dynamo
- Amira / IMOD

Support required

- Single Particle Analysis
- Tomography
- Segmentation/Annotation

Experience

- Beginner/Novice: 83%
- Intermediate: 14%
- Expert: 3%
2017 data generation at SEMC

**EM instrumentation**
- FEI Titan Krios#1 / #2 / #3
- Falcon3 x3
- K2 x3
- FEI Tecnai F20
- DE20
- TVIPS 4K CMOS
- FEI Tecnai Biotwin
- TVIPS 4K CMOS
- JEOL 1230
- Gatan US4000 CCD
- FEI Helios 650
- Quorum cryostage

**Cameras**
- 7 direct detectors on 4 TEMs
- 4 CMOS/CCDs on 3 TEMs + 1 SEM

Since 2015:
- 1,548,411 total images
- (1.1 mil images since 2016)

For 2016:
- 170,933 Krios1 direct detector movies (exposure images)
Data acquisition and processing platforms

**EM instrumentation**

FEI Titan Krios#1 / #2 / #3
Falcon3 x3
K2 x3
FEI Tecnai F20
DE20
TVIPS 4K CMOS

FEI Tecnai Biotwin
TVIPS 4K CMOS

JEOL 1230
Gatan US4000 CCD

FEI Helios 650
Quorum cryostage

Appion

Infrastructure
On the fly data pipeline
c. 2017

Camera -> Buffer server -> File system / cluster

User
Leginon Workstation
Web portal
Data transfer station
Cloud archival storage
SEMC computational support

- HPC Server and storage (DDN):
  - 2 x 42U rack enclosures
  - DDN GRIDScaler GS7K appliance with 1.1PB GPFS paralegal file system
  - 420TB DDN WOS object storage for archival
  - 1056 x CPU cores. 44 x SuperMicro nodes each with 24 x CPU cores and 256GB RAM
  - 4 x GPU nodes each with one GPU and 128GB RAM. One GPU server with 8 x GPUs and 512GB RAM and 2 x GPU servers each with 4 x GPUs and 512GB RAM.
  - 4 buffer servers each with 51TB local storage, 2 x GPUs, 128GB RAM and 10G Fiber Network cards.
  - 5 x 36 QSFP port 56Gb/s FDR InfiniBand switches.
- Bright Cluster Manager
- Basic Onsite Support; 7x24 remote support
Additional computational resources

**Web resources**
- emgweb.nysbc.org
- emg.nysbc.org/redmine
- semc.nysbc.org

**Cluster resources**
- SEMC Goby cluster
- SEMC Guppy cluster
- SEMC cluster

**On site resources**
- **EM session user**
  - Leginon / support workstations (9)
- **Data processing user**
  - Amira workstations (2)
  - GPU workstations (5)
  - Control room workstations (3)
  - Data transfer computers (2)
Example: Single-particle workflow

During EM session

- Setting up workflow
- Data acquisition
- Workflow Optimization if needed
- Frame alignment
- CTF estimation
- Particle picking
- Micrograph/Particle curating

After EM session

- Micrograph/Particle sorting
- Initial 2D classification
- Initial model generation
- 3D refinement

Leginon session

Appion session

SEMC computing

Home institution computing/cluster

2D classification

3D classification

3D refinement

Model building
Example: Single-particle workflow

**During EM session**
- Efficiency
- Example: Single-particle workflow
- Micrograph/Particle sorting
- Initial 2D classification
- Initial model generation
- 3D refinement

**After EM session**
- SEMC computing
- 3D classification
- 3D refinement
- Model building
- Home institution computing/

Efficiency

Example: Single-particle workflow

**During EM session**
- Efficiency
- Example: Single-particle workflow
- Micrograph/Particle sorting
- Initial 2D classification
- Initial model generation
- 3D refinement

**After EM session**
- SEMC computing
- 3D classification
- 3D refinement
- Model building
- Home institution computing/
Web integrated processing

During EM session

Legionon / Appion web interface
Packages offered

After EM session

Appion web interface

CTF Estimation Procedures

During CTF estimation the goal is to fit the standard CTF equation to the power spectra of the electron micrographs.

CTFFIND v4
CTFFIND uses a robust grid search algorithm to find the optimal CTF parameters. Please see the Grigorieff lab website for more information.

GCTF v1.06
This is a GPU accelerated program for real-time CTF determination, refinement, evaluation and correction. Please see the Dynast lab website for more information.

ACE 2
ACE 2 is an unpublished re-implementation of ACE1, but written in objective-C. ACE2 makes several improvements over ACE1 including a several speed enhancements and a robust astigmatism estimate.

Note: It was designed around FEI Tecnai FEG data and other have reported problems using this program.

ACE 1
ACE 1 is the original edge detection program for finding the CTF parameters. Astigmatism estimation never worked quite right in ACE 1 and it has a tendency to give false positives.
Who does the processing workflow?

After EM session

- SEMC computing
  - Micrograph/Particle sorting
  - Initial 2D classification
  - Initial model generation
  - 3D refinement

Home institution computing/

- 2D classification
- 3D classification
- 3D refinement
- Model building

Appion web interface

Cluster Stack List

- 160 Class Averages [variances] (ID 41)
- 80 Class Averages [variances] (ID 40)
Who does the processing workflow?

After EM session

**SEMCC computing**
- Micrograph/Particle sorting
- Initial 2D classification
- Initial model generation
- 3D refinement

**Home institution computing**
- 2D classification
- 3D classification
- 3D refinement
- Model building

Appion web interface for TEM use

Cluster Stack List

- Clustering info: bender3 (DD: 12) with 1 clusters
  - Type: Xnipp_KerDen SDF
- Clustering info: center1 (DD: 12) with 2 clusters
  - Type: SPIDER Cenet
    - Method: Factor list
      - 160 Class Averages (variance) (DD 41)
      - 80 Class Averages (variance) (DD 40)

- Clustering info: bender1 (DD: 11) with 1 clusters
  - Type: Xnipp_KerDen SDF
Too busy to improve?
Feedback from the community

Development

Production

Time / Resources

Strategy
Processing at SEMC

**Community**
What is our user community asking for?

**Infrastructure**
What is our infrastructure?

**Efficiency**
How does processing lead to more efficient data collection?

**Balance**
What is the balance between staff and user effort?

**Strategy**
How do we evolve and prepare for future advances?
PROCESSING (cont)

before scaling up and support issues

...Anchi
Before 3 Krioses

**EM instrumentation**
- FEI Titan Krios#1
  - K2 x1
- FEI Tecnai F20
  - DE20
- TVIPS 4K CMOS
- JEOL 1230
- Gatan US4000 CCD
- FEI Helios 650
- Quorum cryostage

**Cameras**
- 2 direct detectors on 2 TEMs
- 3 CMOS/CCDs on 1 TEMs + 1 SEM
2015-2016 computation support

- Guppy (8 nodes with 4-8 CPU cores)
- 1 Database and 1 web server
- 3 gpu workstations
- Individually mounted storage 300 TB

- GPFS storage 0.5 PB
- DE20 dedicated processing server (24 Processors)
- 1056 x CPU cores. 44 x SuperMicro nodes each with 24 x CPU cores and 256GB RAM
- 4 x GPU nodes each with one GPU and 128GB RAM.
Typical Problems

During EM session

- Frame alignment from the last user is not finished
- Web service image loading can not keep up with multiple user viewing.
- Database query lock up.
- Computation resource upgrade in progress

After EM session

- Inefficient request of resource
  - Ask more than needed
  - Ask less than required
- Small one processor jobs compete with MPI processes.
- Users do not (know how to) clean up bad processing results.
- New development breaks the pipeline.