Automated Data Integration at NE-CAT

The RAPD Integration Pipeline

David B. Neau

Cornell University / NE-CAT

Initial goal of integration pipeline

- Provide users a quick initial feedback on the quality of their data.
    - Is the data good enough to use?
    - Initial estimate of resolution
    - Can I move on to other projects, or shift my goal from collecting data that work to collecting data that will work better?
- Not viewed as a replacement for manual processing of data.

Current goal of pipeline

- Provide integrated and scaled data suitable for use in structure solution and/or refinement for all but the most difficult cases.
  - Process appropriately any data that an experienced crystallographer would be able to easily process.

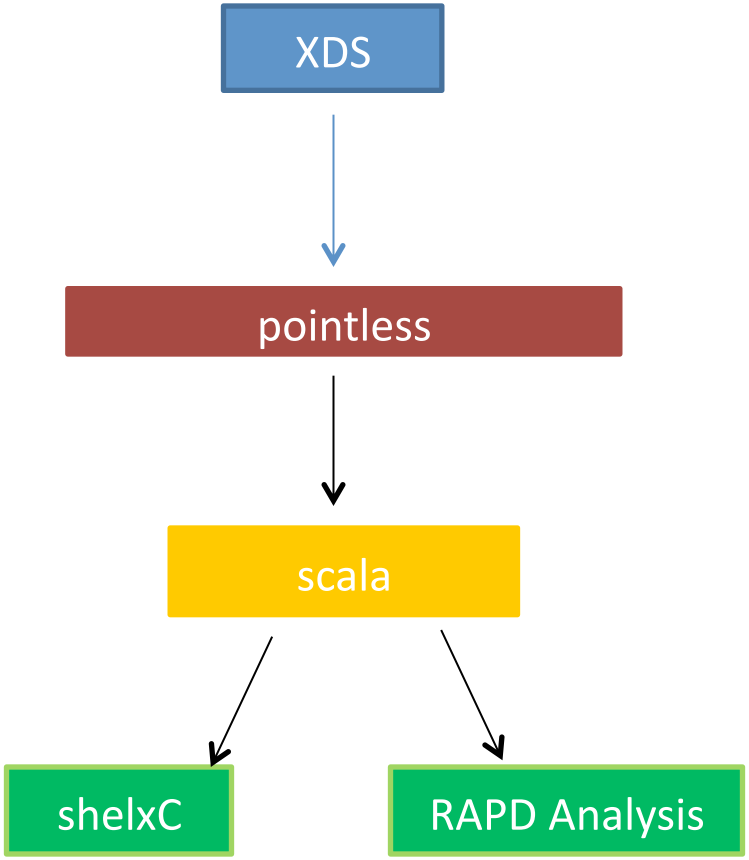
Current capability of pipeline

- 24,543 integrations as of July 24, 2010
- Successfully processes "normal" data at moderate to high resolution

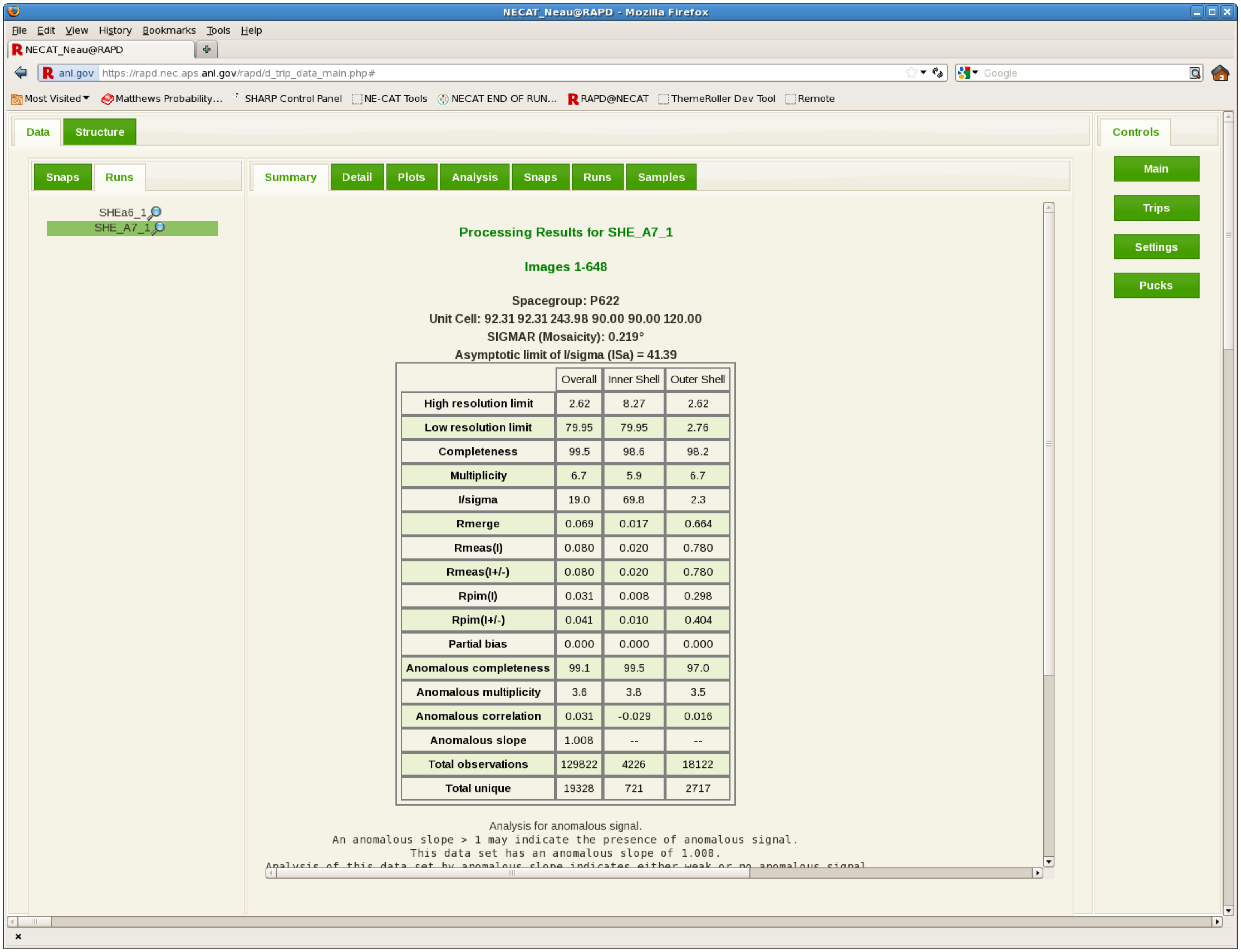
- Struggles at times with low resolution data.
- Does not account for pathologies such as twinning, multiple lattices, anisotropy.

Pipeline Layout

- XDS
- Pointless
- Scala
- The above steps may be launched multiple times during a data collection
- End result is processed data in multiple file formats, along with some log files and input files, all of which can be downloaded by the user.
- shelxC if anomalous slope from scala is greater than 1
- RAPD Analysis pipeline – Jon Schuermann



RAPD Integration



Processing Results for SHE_A7_1

Images 1-648

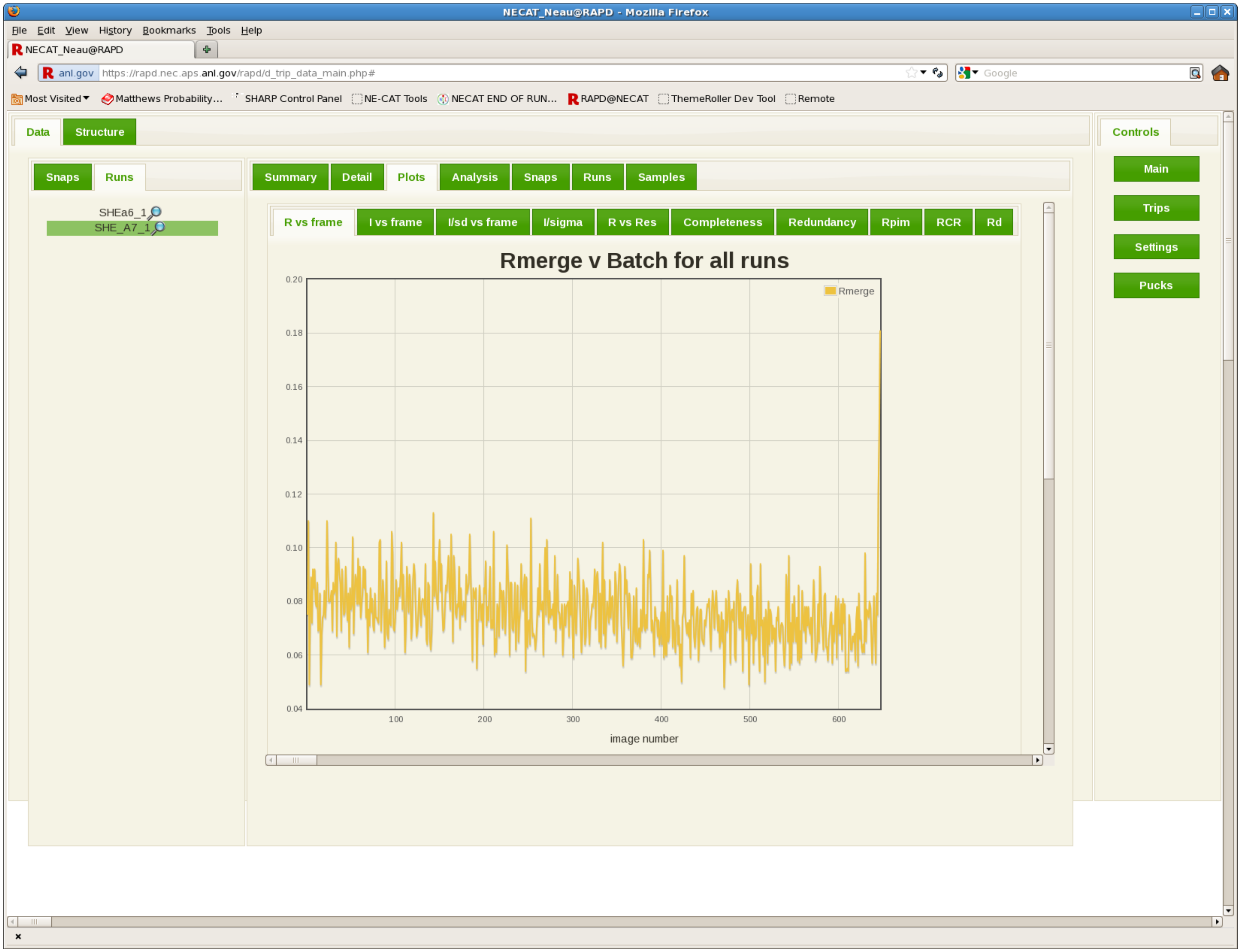
Spacegroup: P622
Unit Cell: 92.31 92.31 243.98 90.00 90.00 120.00
SIGMAR (Mosaicity): 0.219°
Asymptotic limit of I/sigma (ISa) = 41.39

	Overall	Inner Shell	Outer Shell
High resolution limit	2.62	8.27	2.62
Low resolution limit	79.95	79.95	2.76
Completeness	99.5	98.6	98.2
Multiplicity	6.7	5.9	6.7
I/sigma	19.0	69.8	2.3
Rmerge	0.069	0.017	0.664
Rmeas(I)	0.080	0.020	0.780
Rmeas(I+/-)	0.080	0.020	0.780
Rpim(I)	0.031	0.008	0.298
Rpim(I+/-)	0.041	0.010	0.404
Partial bias	0.000	0.000	0.000
Anomalous completeness	99.1	99.5	97.0
Anomalous multiplicity	3.6	3.8	3.5
Anomalous correlation	0.031	-0.029	0.016
Anomalous slope	1.008	--	--
Total observations	129822	4226	18122
Total unique	19328	721	2717

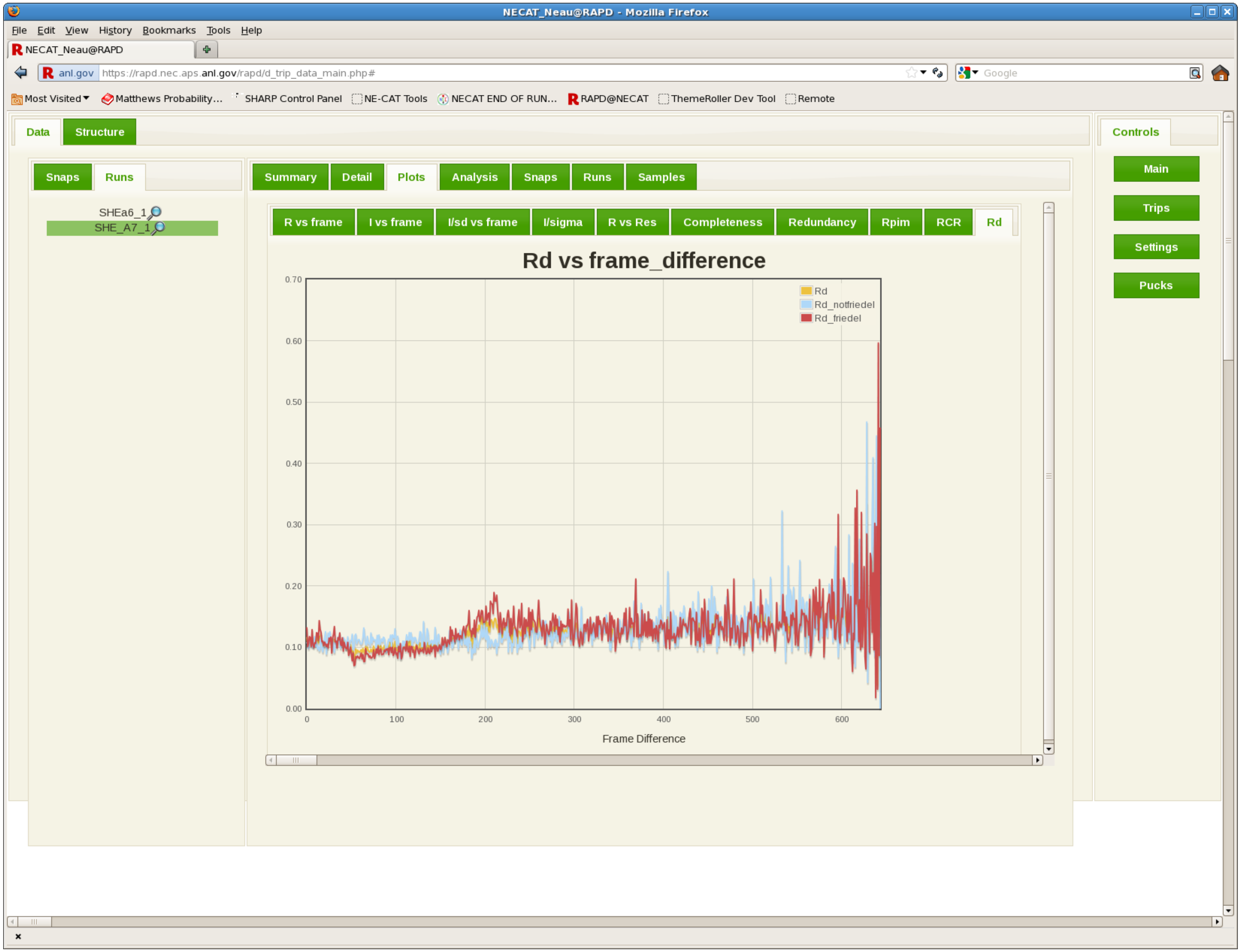
Analysis for anomalous signal.

```
An anomalous slope > 1 may indicate the presence of anomalous signal.
This data set has an anomalous slope of 1.008.
Analysis of this data set by anomalous slope indicates either weak or no anomalous signal.
```

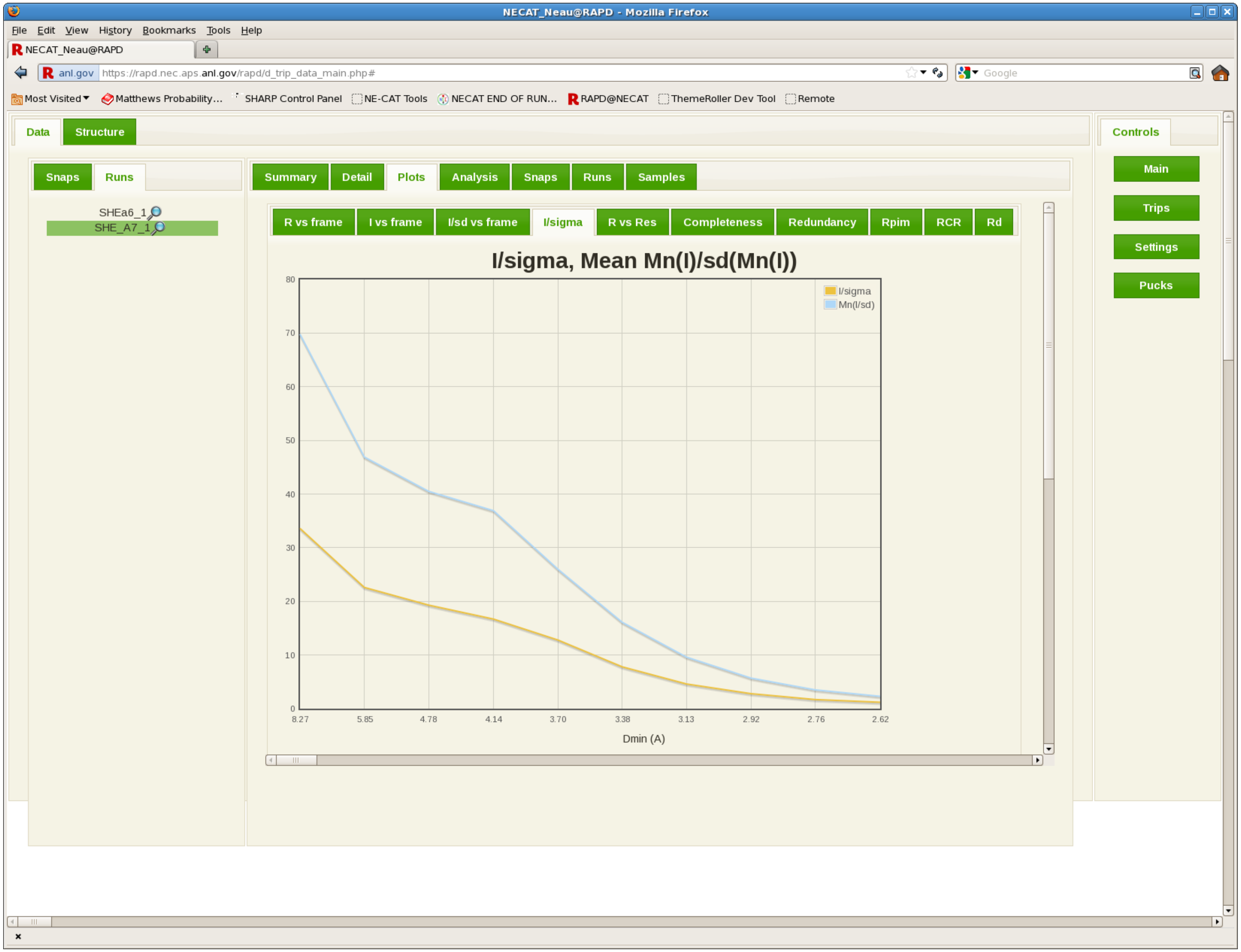
RAPD Integration



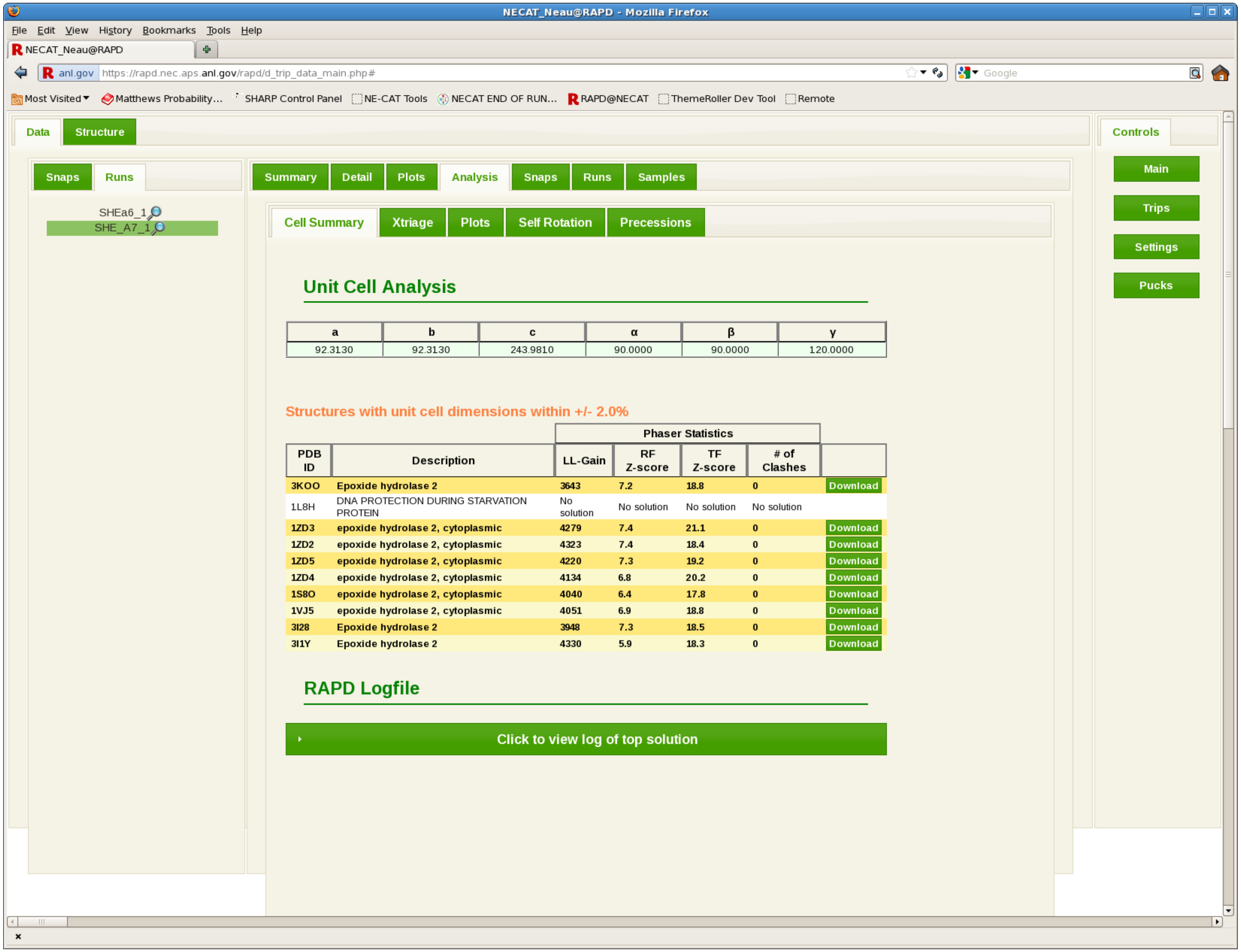
RAPD Integration



RAPD Integration



RAPD Integration



Unit Cell Analysis

a	b	c	α	β	γ
92.3130	92.3130	243.9810	90.0000	90.0000	120.0000

Structures with unit cell dimensions within +/- 2.0%

		Phaser Statistics				
PDB ID	Description	LL-Gain	RF Z-score	TF Z-score	# of Clashes	
3KOO	Epoxide hydrolase 2	3643	7.2	18.8	0	Download
1L8H	DNA PROTECTION DURING STARVATION PROTEIN	No solution	No solution	No solution	No solution	
1ZD3	epoxide hydrolase 2, cytoplasmic	4279	7.4	21.1	0	Download
1ZD2	epoxide hydrolase 2, cytoplasmic	4323	7.4	18.4	0	Download
1ZD5	epoxide hydrolase 2, cytoplasmic	4220	7.3	19.2	0	Download
1ZD4	epoxide hydrolase 2, cytoplasmic	4134	6.8	20.2	0	Download
1S8O	epoxide hydrolase 2, cytoplasmic	4040	6.4	17.8	0	Download
1VJ5	epoxide hydrolase 2, cytoplasmic	4051	6.9	18.8	0	Download
3I28	Epoxide hydrolase 2	3948	7.3	18.5	0	Download
3I1Y	Epoxide hydrolase 2	4330	5.9	18.3	0	Download

RAPD Logfile

Speed of Pipeline

Data collected on ADSC in binned mode

95 data frames, 1 sec exposure, 1° oscillation, 1.95Å resolution
$P2_12_12_1$; a = 75Å, b = 121Å, c = 131Å

Total data collection time: 6 min 26 sec
Time from last frame
to initial integration results: 1 min 17 seconds
to final integration results: 2 min 20 seconds
to pipeline completion: 4 min 53 seconds
(Analysis found 10 unit cell matches in PDB)
Intermediate integration results at 10, 20, 30, 40, 60, and 80 frames

Speed of Pipeline

Data collected on Pilatus

120 data frames, 1 sec exposure, 0.1° oscillation, 3.54Å resolution
P222 ; a = 203Å, b = 450Å, c = 622Å

Total data collection time: 2 min

Time from last frame

to initial integration results: 1 min 28 sec

to final integration results: 2 min 59 sec

to pipeline completion: 20 min

(Analysis found 10 unit cell matches in PDB)

Intermediate integration results at 40 frames

Speed of Pipeline

Data collected on Pilatus

600 data frames, 0.1 sec exposure, 0.1° oscillation, 2.94Å resolution
$P6_122$; a = 92Å, c = 243Å

Total data collection time: 1 min

Time from last frame

- to initial integration results: 1 min 5 sec
- to final integration results: 2 min 2 sec
- to pipeline completion: 4 min 43 sec
  - (Analysis found 10 unit cell matches in PDB)

No intermediate results

Reintegration pipeline

- Similar data flow to integration
- Allows user to select a subset of a data set by choosing a new start and end frame.
- Most often used when a user suspects significant radiation damage has occurred.
- Future plans
  - make pipeline identical to integration
  - Allow user to include their own desired parameters.

Merging pipeline

- Allows user to choose two data sets and merge them.
- Pipeline takes the mtz files generated by pointless and uses scala to merge the files.
- Users may bootstrap multiple data sets together.
- Future plans
  - Multiple data set merging
  - “Smart” merging

Future Goals/Wishes of RAPD

- Speed… much more speed (threading detector modules)
- XDS mosaicity calculated from small wedges of ‘snapshots’
- Shift libraries to CCTBX for compatibility
- More database mining
- Run auto-beam center calculation in background
- MAD and MRSAD pipelines
- New GUI
- Better hooks for wrapping
- Multiple lattice integration
- Better support for multi-wedge data collection

Acknowledgments

- Frank Murphy – Core and user interface
- David Neau – data (re)processing
- Kay Perry, Surajit Banerjee

- RAPD
  - https://rapd.nec.aps.anl.gov/rapd
- Wiki
  - https://rapd.nec.aps.anl.gov/wiki/main_page

Where we’d like to go

Faster completion of pipeline

Fast enough the it’s a no-brainer for the user to wait for results before moving on to a new crystal (30 seconds or less?)

As much parallelization as possible

Fairly simple for beamlines to be a center of computing resources

Elimination of having to edit text scripts as much as possible

Ideal would be having an integration service running which only requires passing the experiment details to it.

One that can start to process data even if all of the frames are not there yet.

More robust integration

More decision making information

What does the user need to know and how to present it

What can the program know, and perhaps launch other pipelines based on the data results

i.e. if anomalous signal is detected – launch a SAD pipeline

Knowledge of what the user is seeking and how best to meet those needs

Knowledge of previously collected data sets and/or future data sets

Ability for users to access pipeline (and potentially computing resources) offline.