Yale/NIDA Neuroproteomics Center

Biostatistics and Bioinformatics Core
Director: Kei Cheung, Ph.D.
Associate Director: Perry Miller, M.D., Ph.D.
Yale Center for Medical Informatics
BBC is divided into four sections:

- High Performance Computing (Nick Carriero and Rob Bjornson)
- Biostatistics data analysis (Hongyu Zhao and Lisa Chung)
- Bioinformatics (Mark Gerstein and Can Bruce)
- Yale Protein Expression Database (Kei Cheung, Perry Miller, and Mark Shifman)

Regular bi-weekly meetings:

- Plan and keep track of BBC activities
- Coordination between BBC and other cores
- Invited talks
The HPC infrastructure is supported by the Center for HPC in Biology and Biomedicine

- The HPC center currently has two clusters: BulldogN (1,536 cores) and Louise (3,464 cores) with a total storage of 1.94 PB. Louise has been used by >400 researchers from 107 research groups and 33 departments during 2010-2012.

- X!!Tandem: Parallelization of the popular X!Tandem Mass Spectrometry tool to speed up identification of proteins.

- Peptide uniqueness: Two large runs (totaling ~1,000,000 peptides) of a pipeline that compares a peptide for uniqueness against up to date versions of standard references (SwissProt, TREMBL).
  - Results are integrated with YPED so that they can be used to select peptides that are specific to a particular protein target.

- Data conversion: Converting MS data from a vendor specific output format to a more generic format (mzML).
Phosphorylation and Motifs: Development of a tool to process information about peptides and candidate phosphorylation sites into expanded amino acid sequence data in a format suitable for motif detection.

Exploratory work on deploying OpenMS/OpenSWATH (http://open-ms.sourceforge.net/) on our HPC cluster.

Other activities include routine maintenance, bug fixes, recovery work, generalizing input and output handling, and test runs to assess broader applicability to NIDA groups.
Implementing R applications for data visualization and exploratory analysis

- Biostatistics analysis for the following:
  - Multiple Reaction Monitoring (MRM)
  - SWATH
  - iTRAQ experiments
R applications: getMRMplot and getSWATHplot

- Support MRM and SWATH output from Multiquant (AB Sciex) and Skyline (MacCoss Lab).

- Provide graphical presentation of the following:
  - peak quality assessment using peak area, signal-to-noise ratio, and retention time
  - reproducibility analysis among replicated samples
  - sample quality assessment
  - peak area visualization across multiple samples
  - data normalization
Data Visualization and Exploratory Analysis

**Signal/Noise Quality**

- Great (cat SN = 5): 181 (53.6%)
- Good (cat SN = 4): 107 (31.7%)
- Detect (3 < cat SN < 4): 37 (10.6%)
- Noise (cat SN < 3): 13 (3.8%)

**Sample-to-Sample Correlation**

**Reproducibility**

**Peak Area Normalization**
Data Normalization

Quality Assessment and Exploratory Analysis

Assess peak/transition/sample quality, data filtering, clustering analysis, reproducibility analysis

To minimize the effects due to:
- small difference in protein quantities
- fluctuations generated by technique or experimental protocol

approaches:
- median adjustment
- quantile normalization
- rank invariant

Identification of differentially expressed proteins/peptides

Peptide- and protein-level fold change and significance analysis incorporating experimental design structure
- t-test for pairwise comparison
- linear model and mixed effect model
- moderated t- and F-test (Smyth 2004)
analyses of mRNA & protein abundances in yeast
overall correlation 66%
increased correlation in some subsets defined by subcellular localization or functional groups
PARE: tool automating these analyses
- This can be used by NIDA researchers to correlate subsets of genes with neuro-related functions
Bioinformatics Core Activities

Analyze list of differentially expressed proteins from proteomics experiments

- Establish thresholds, identify outliers.
- Determine biological processes, cellular components, biochemical functions and other database annotation terms that are significantly enriched.
- Identify targets of transcription factors, kinases that are significantly enriched.
- Identify likely pairs of interacting proteins.
- Generate likely interaction networks to develop hypotheses regarding connections between perturbed protein and observed expression changes.
NIDA Projects

- **S. Sathyanesan**

- **S. Chandra**

- **M. Morabito**
  - Effect of MDM2 expression on mouse brain proteome in a P53⁻/⁻ background.
  - Time course of proteome changes following NMDA treatment.
  - Effect of Roscovitin, a CDK5 inhibitor, on brain proteome.
  - Effect of knocking out P35, a CDK5 activator, on brain proteome.

- **A. Stipanovich / A. Nairn**
  - Casein Kinase Delta substrates in mouse brain, studied by SILAM
YPED is a comprehensive suite of tools designed to cover a broad spectrum of techniques for quantitative proteomics (discovery and targeted proteomics; and labeled and label-free quantitation).

It captures data produced by a wide range of MS instruments and technologies, and presents them via the Web as a set of relevant results that are understandable for non-specialists.

It implements several data access privileges based on different user types including core lab users, researchers (PIs and their lab members), anonymous reviewers and public users.

The database is implemented using Oracle with a web front (Java).
As of April 2013, YPED is being used by 1313 researchers from 551 principal investigators at 325 institutions. There are 143 NIDA users.

It contains 15,142 datasets resulting in a spectral library which encompasses 626,695 distinct proteins and 3,008,435 distinct peptides.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Blast Protein Count*</th>
<th>Blast Peptide Count*</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.Coli</td>
<td>3,970</td>
<td>41,760</td>
</tr>
<tr>
<td>Yeast</td>
<td>5,684</td>
<td>54,948</td>
</tr>
<tr>
<td>Rat</td>
<td>11,962</td>
<td>122,322</td>
</tr>
<tr>
<td>Mouse</td>
<td>20,059</td>
<td>214,905</td>
</tr>
<tr>
<td>Human</td>
<td>20,843</td>
<td>243,749</td>
</tr>
</tbody>
</table>
## YPED User Interface

### PI Functions User: Angus Nairn

66 projects found, displaying 1 to 10. [First/Prev] 1, 2, 3, 4, 5, 6, 7 [Next/Last]

<table>
<thead>
<tr>
<th>Project Name</th>
<th>Samples</th>
<th>Analysis Types</th>
<th>Organisms</th>
<th>PI</th>
<th>Date Created</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 CEX fractions Human brain on 5G00T1</td>
<td>17</td>
<td>LCMS</td>
<td>Homo sapiens</td>
<td>Angus Nairn</td>
<td>2012-07-31 16:14:34.0</td>
</tr>
<tr>
<td>5G00TT PPA2 Methyl-leucine searches 082211</td>
<td>10</td>
<td>LCMS</td>
<td>Homo sapiens, Mus musculus</td>
<td>Angus Nairn</td>
<td>2011-06-22 09:21:20.0</td>
</tr>
<tr>
<td>5G00TT Ret PSD - 50 min runs</td>
<td>1</td>
<td>LCMS</td>
<td>Rattus norvegicus</td>
<td>Angus Nairn</td>
<td>2012-01-31 13:29:26.0</td>
</tr>
<tr>
<td>5G00TT methylation on PPA2</td>
<td>3</td>
<td>LCMS</td>
<td>Bos taurus</td>
<td>Angus Nairn</td>
<td>2012-04-24 10:46:41.0</td>
</tr>
<tr>
<td>ACID EX</td>
<td>8</td>
<td>LCMS, LabelFreeQuant</td>
<td>Rattus norvegicus</td>
<td>Angus Nairn</td>
<td>2012-10-26 17:30:52.0</td>
</tr>
<tr>
<td>B2 and C8 sub-cellular fractionation</td>
<td>10</td>
<td>LCMS</td>
<td>Homo sapiens</td>
<td>Angus Nairn</td>
<td>2012-07-31 16:04:24.0</td>
</tr>
<tr>
<td>CONT OKA_CSA Enr/F on 5G00TT</td>
<td>6</td>
<td>LCMS</td>
<td>Homo sapiens</td>
<td>Angus Nairn</td>
<td>2011-06-21 11:36:53.0</td>
</tr>
<tr>
<td>Cocaine PSD Mouse ITRAQ</td>
<td>14</td>
<td>ITRAQ8plex, LCMS</td>
<td>Rattus norvegicus</td>
<td>Angus Nairn</td>
<td>2008-11-04 16:42:42.0</td>
</tr>
<tr>
<td>Comprised SILAC for PF-Tag</td>
<td>14</td>
<td>LCMS</td>
<td>Homo sapiens</td>
<td>Angus Nairn</td>
<td>2011-03-31 11:38:30.0</td>
</tr>
<tr>
<td>DARPP32, JPH1 intim</td>
<td>8</td>
<td>LCMS</td>
<td>Rattus norvegicus</td>
<td>Angus Nairn</td>
<td>2011-06-27 15:32:18.0</td>
</tr>
</tbody>
</table>

1,393 results found, displaying 181 to 190. [First/Prev] 15, 16, 17, 18, 19, 20, 21, 22 [Next/Last]

<table>
<thead>
<tr>
<th>Analysis Type</th>
<th>Sample Name</th>
<th>Search Engine</th>
<th>Database</th>
<th>Date</th>
<th>User/PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCMS</td>
<td>Sakabe_Cortex-PSD_120511_90minID0A_1_500_3_info</td>
<td>MASCOT</td>
<td>1:contaminants_20090624 fasta 2:SwissProt_2012_05 fasta</td>
<td>2012-06-05 09:13:05.0</td>
<td>Fumiko Sakabe/Angus Nairn</td>
</tr>
<tr>
<td>LCMS</td>
<td>WTI#2 C8C En info</td>
<td>MASCOT</td>
<td>SwissProt_2012_05 fasta</td>
<td>2012-06-04 12:46:40.0</td>
<td>Veronica Musante/Angus Nairn</td>
</tr>
<tr>
<td>LCMS</td>
<td>WTI#2 C8C En info</td>
<td>MASCOT</td>
<td>SwissProt_2012_05 fasta</td>
<td>2012-06-04 12:42:54.0</td>
<td>Veronica Musante/Angus Nairn</td>
</tr>
<tr>
<td>LCMS</td>
<td>WTI#1 C8C En info</td>
<td>MASCOT</td>
<td>SwissProt_2012_05 fasta</td>
<td>2012-06-04 12:35:38.0</td>
<td>Veronica Musante/Angus Nairn</td>
</tr>
<tr>
<td>LCMS</td>
<td>WTI#1 C8C En info</td>
<td>MASCOT</td>
<td>SwissProt_2012_05 fasta</td>
<td>2012-06-04 12:30:26.0</td>
<td>Veronica Musante/Angus Nairn</td>
</tr>
<tr>
<td>LCMS</td>
<td>HEK B6 20 CEX fractions Elite Info</td>
<td>MASCOT</td>
<td>SwissProt_2012_05 fasta</td>
<td>2012-06-01 15:04:37.0</td>
<td>Robert Kitchen/Angus Nairn</td>
</tr>
<tr>
<td>LCMS</td>
<td>HEK B6 19 CEX fractions Elite info</td>
<td>MASCOT</td>
<td>SwissProt_2012_04 fasta</td>
<td>2012-06-01 09:32:14.0</td>
<td>Robert Kitchen/Angus Nairn</td>
</tr>
<tr>
<td>LCMS</td>
<td>Kitchen Human Cerebellum, Orbitrap Elite, 20 CEX fractions info</td>
<td>MASCOT</td>
<td>SwissProt_2012_04 fasta</td>
<td>2012-06-01 17:52:10.0</td>
<td>Robert Kitchen/Angus Nairn</td>
</tr>
<tr>
<td>LCMS</td>
<td>PSD JPT.com heavy peptides JPT database info</td>
<td>MASCOT</td>
<td>SwissProt_2012_04 fasta</td>
<td>2012-05-10 11:46:25.0</td>
<td>Fumiko Sakabe/Angus Nairn</td>
</tr>
<tr>
<td>LCMS</td>
<td>PSD JPT.com heavy peptides MOUSE/RAT search info</td>
<td>MASCOT</td>
<td>SwissProt_2012_04 fasta</td>
<td>2012-05-03 10:59:01.0</td>
<td>Fumiko Sakabe/Angus Nairn</td>
</tr>
</tbody>
</table>

For questions or comments contact [Mark Shifman](mailto:mark.shifman@yale.edu)

Updated 22 Apr 2013 16:44
YPED User Interface (Cont’d)
YPED Repository

- It contains the results of projects which have been released for public viewing by the PI’s along with raw data from the samples.

- It provides an access code provision for viewing results prior to public release. This feature is useful for making the results available to reviewers and collaborators who do not have YPED access.
Acknowledgement

- Keck Biotechnology Resource Laboratory at Yale
  - Christopher Colangelo
  - Erol Gulcicek
  - Tukiet Lam
  - Kathy Stone
  - Kenneth Williams
  - Terence Wu

- Department of Psychiatry, Yale School of Medicine
  - Angus Nairn

- Nairn Lab and Gerstein Lab
  - Robert Kitchen