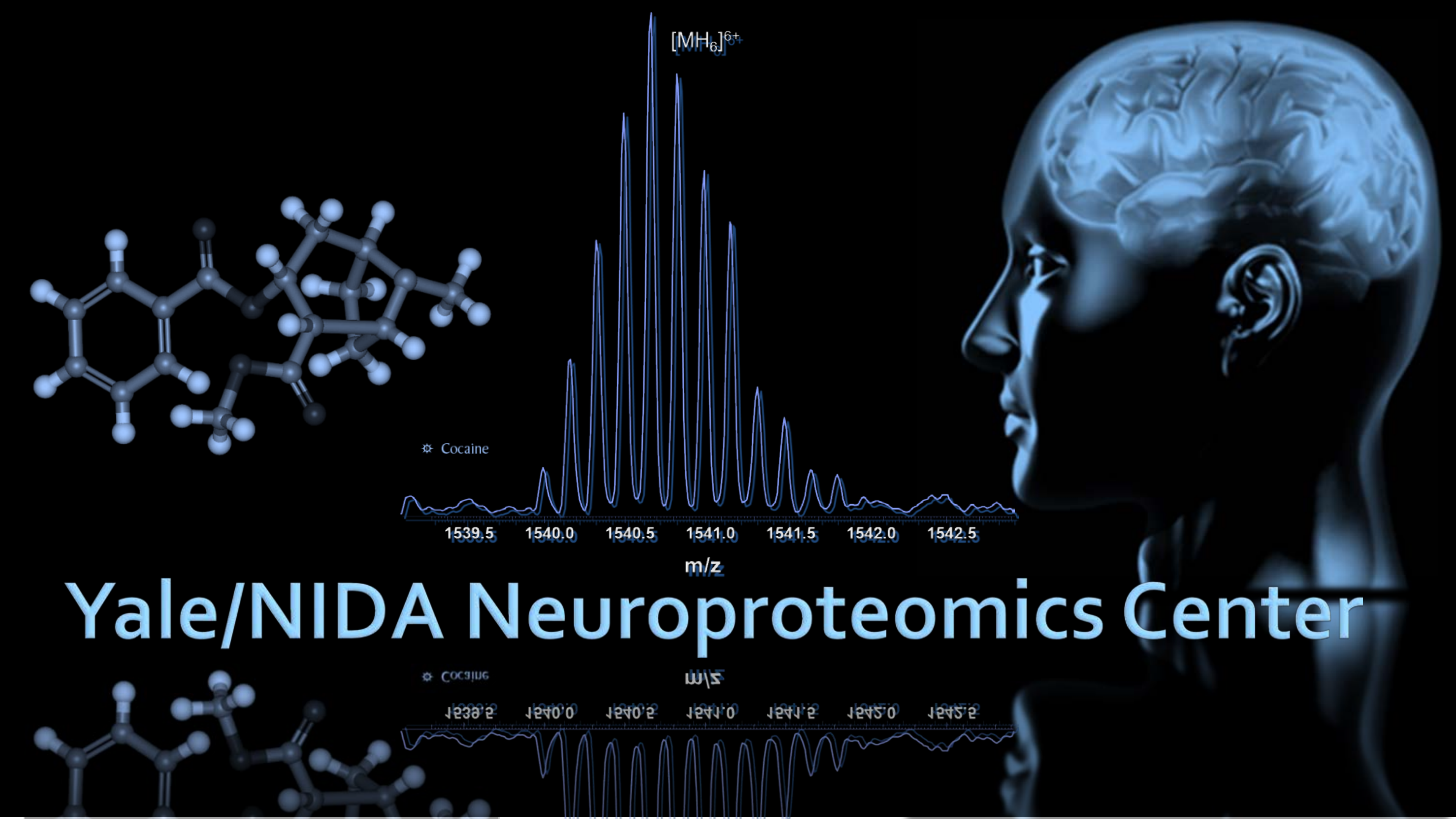


Yale/NIDA Neuroproteomics Center

Biostatistics and Bioinformatics Core

Director: Kei Cheung, Ph.D.

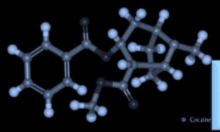
Associate Director: Perry Miller, M.D., Ph.D.

Yale Center for Medical Informatics

BBC Structure and Management

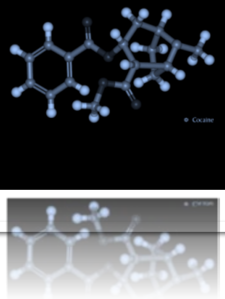
- BBC is divided into four sections:
  - High Performance Computing (Nick Carriero and Rob Bjornson)
  - Biostatistics data analysis (Hongyu Zhao and Lisa Chung)
  - Bioinformatics (Mark Gerstein and Can Bruce)
  - Yale Protein Expression Database (Kei Cheung, Perry Miller, and Mark Shifman)

- Regular bi-weekly meetings:
  - Plan and keep track of BBC activities
  - Coordination between BBC and other cores
  - Invited talks



HPC Infrastructure and Activities

- 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).

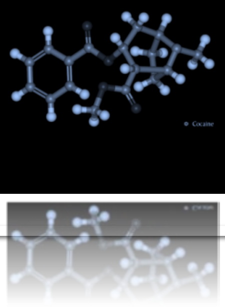


HPC Activities

- 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.



Biostatistics Section

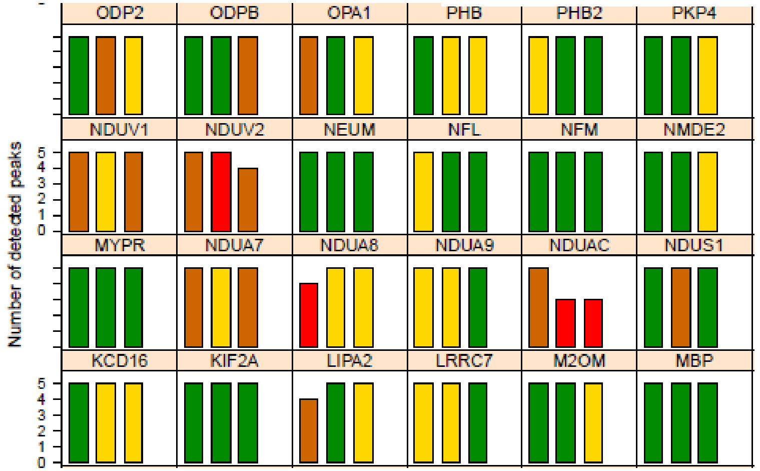
- Implementing R applications for data visualization and exploratory analysis
- Biostatistics analysis for the following:
  - Multiple Reaction Monitoring (MRM)
  - SWATH
  - iTRAQ experiments

R applications: getMRMplot and getSWATHplot

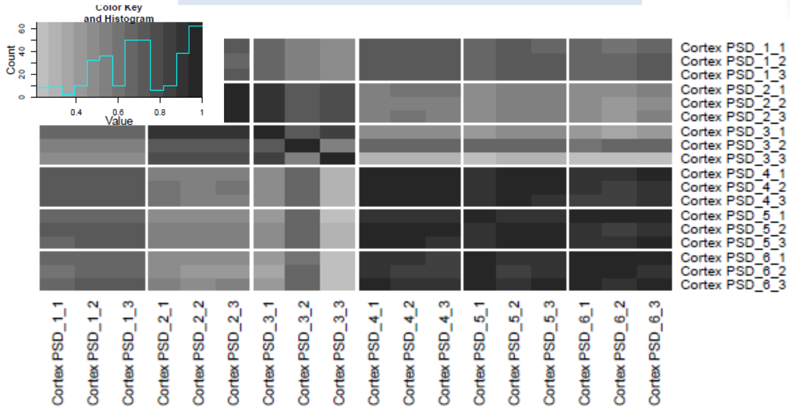
- Support MRM and SWATH output from Multiquant (AB Sciex) and Skyline (MacCoss Lab).
- Provide graphical presentation of the following:
  - peak quality assessment using peak area, signal-to-noise ratio, and retention time
  - reproducibility analysis among replicated samples
  - sample quality assessment
  - peak area visualization across multiple samples
  - data normalization

Data Visualization and Exploratory Analysis

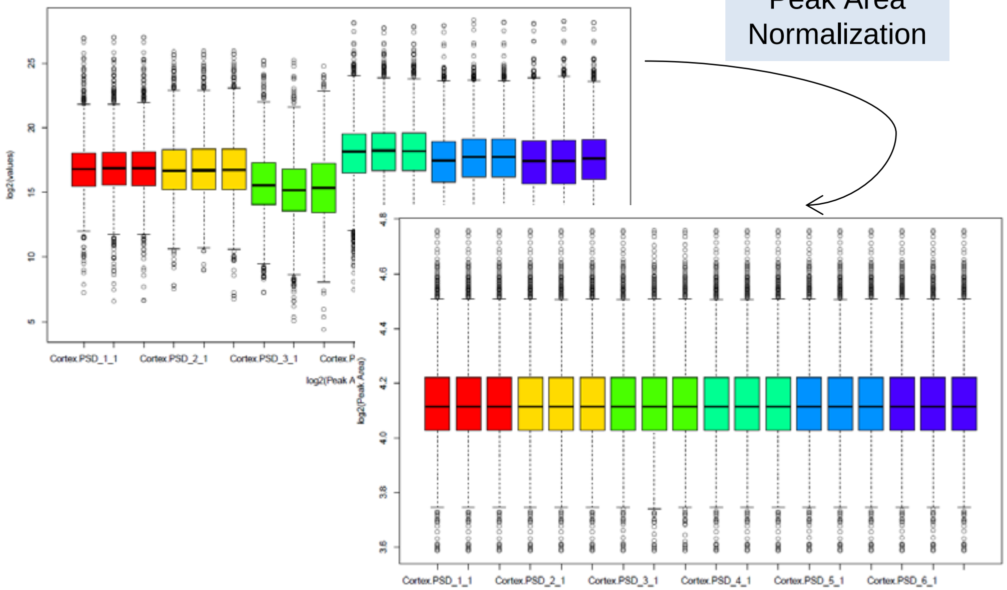
Signal/Noise Quality



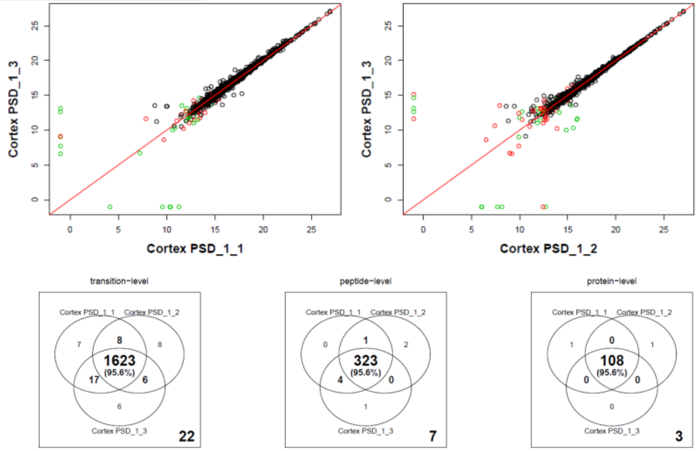
Sample-to-Sample Correlation



Peak Area Normalization

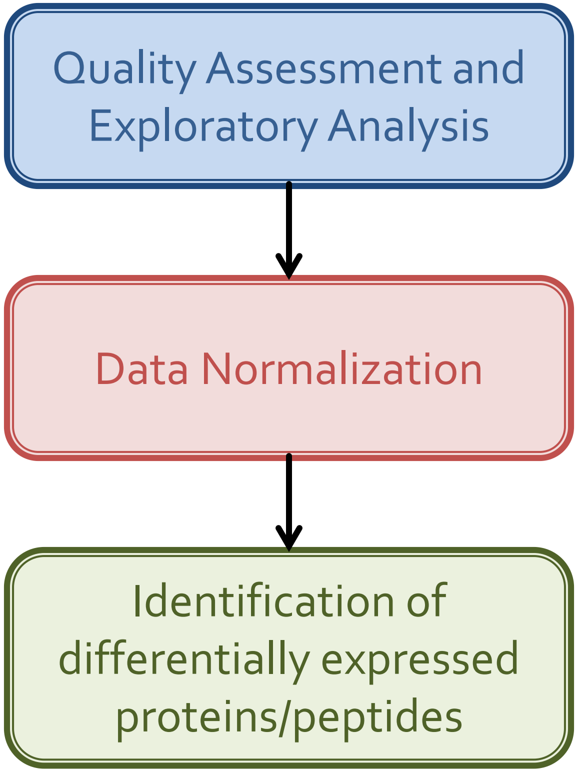


Reproducibility



Yale NIDA Proteomics Center

Biostatistics Analysis



Quality Assessment and Exploratory Analysis

Data Normalization

Identification of differentially expressed proteins/peptides

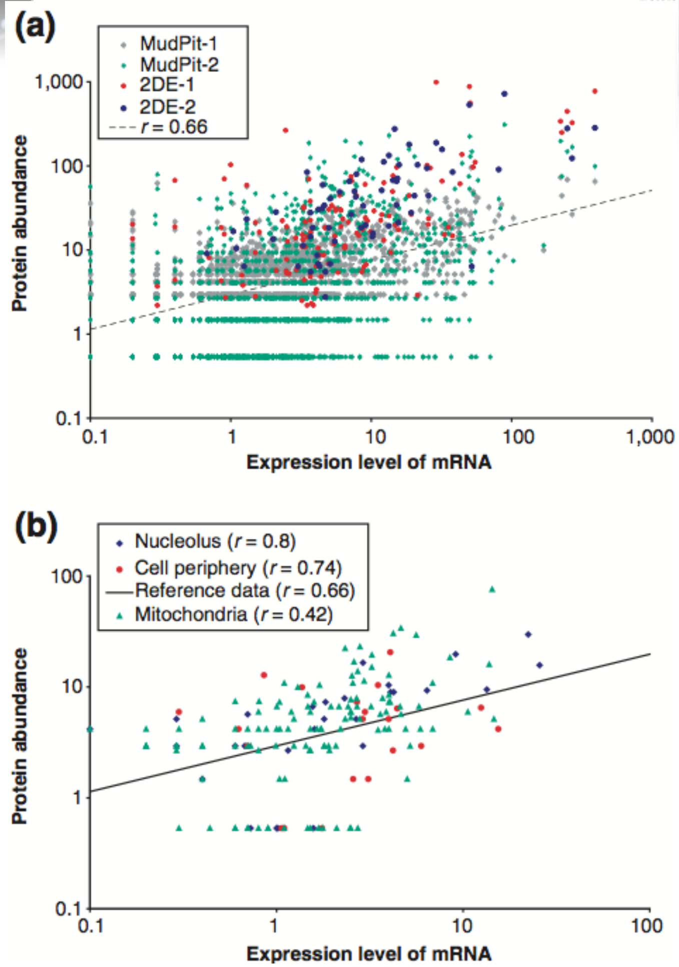
Assess peak/transition/sample quality, data filtering, clustering analysis, reproducibility analysis

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

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)

Bioinformatics: mRNA / Protein Correlation



Opinion

Comparing protein abundance and mRNA expression levels on a genomic scale

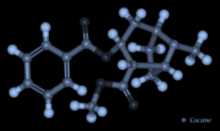
Dov Greenbaum*, Christopher Colangelo†‡, Kenneth Williams†‡ and Mark Gerstein†§

- 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**
  - Sathyanesan, M., Girgenti, M.J., Banasr, M., Stone, K., Bruce, C., Guilchicek, E., Wilczak-Havill, K., Nairn, A., Williams, K., Sass, S., Duman, J.G., Newton, S.S. (2012) A molecular characterization of the choroid plexus and stress-induced gene regulation. Transl Psychiatry 2(7): e139.
- **S. Chandra**
  - Zhang,Y., Henderson, M.X., Colangelo, C.M., Ginsberg, S., Bruce, C., Wu, T., and Chandra, S.S. (2012) Identification of CSPα clients reveals a role in dynamin 1 regulation. Neuron 74(1):136-50.
- **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

Yale Protein Expression Database

- 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).

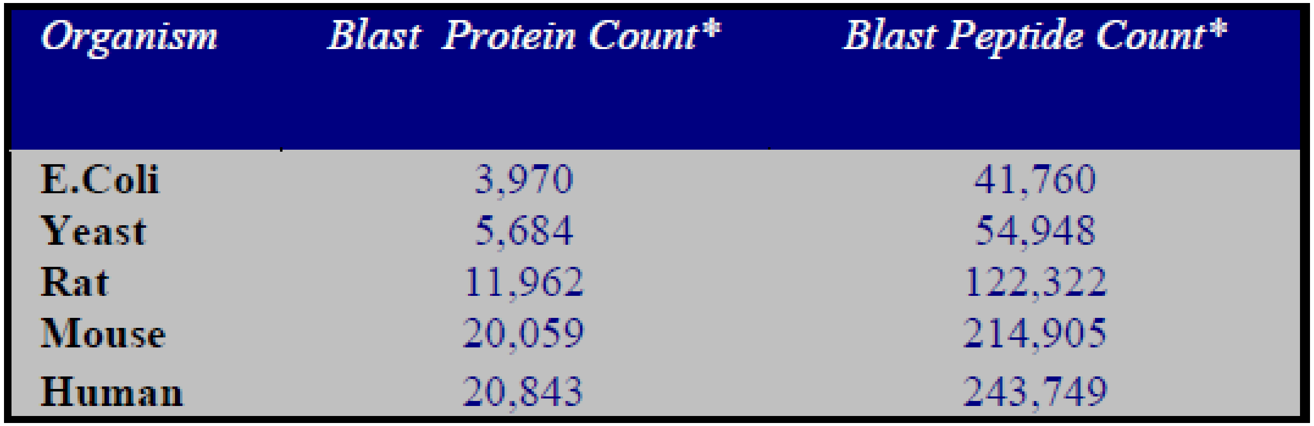
- It captures data produced by a wide range of MS instruments and technologies, and presents them via the Web as a set of relevant results that are understandable for non-specialists.

- 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).

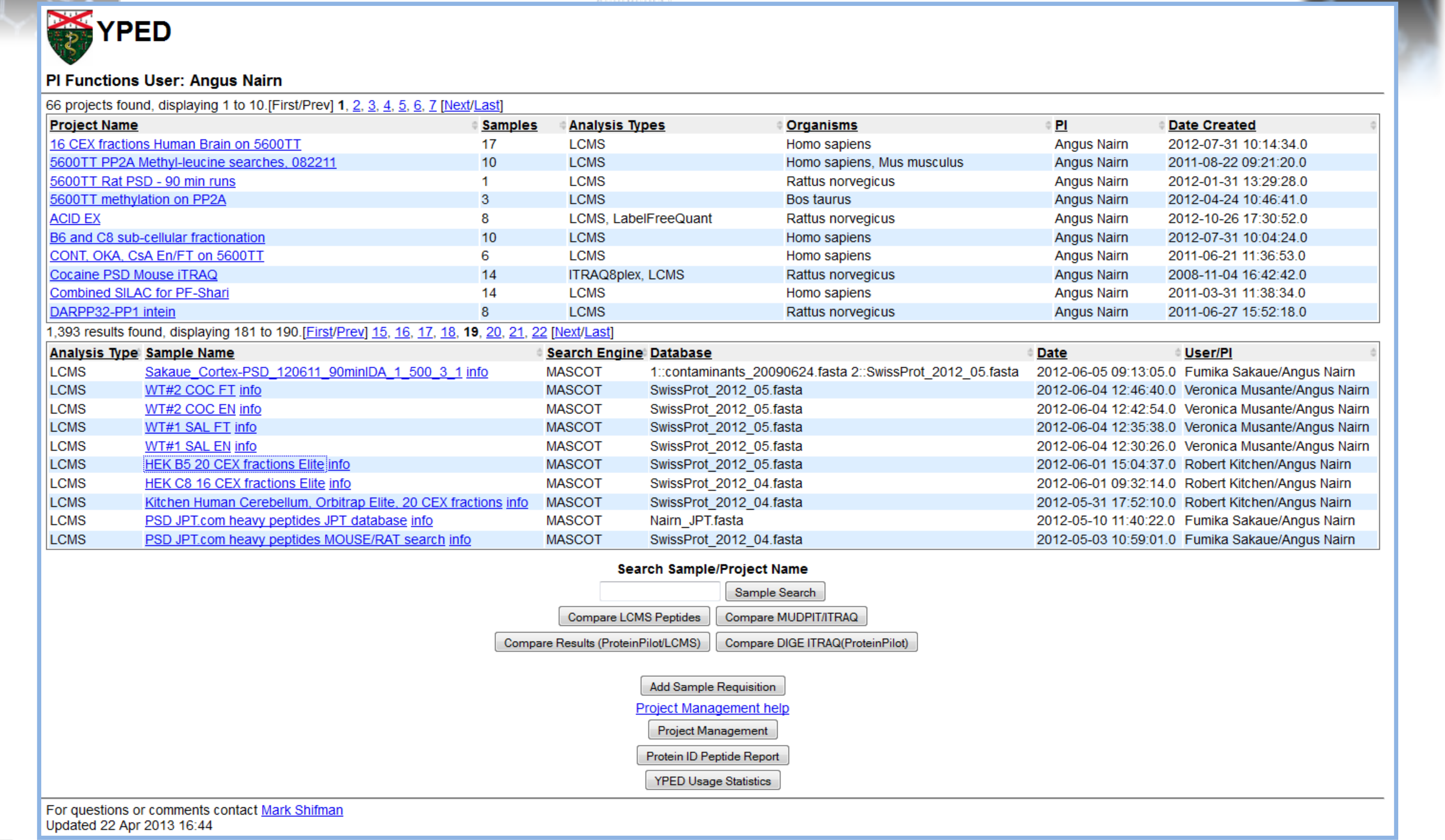
YPED Usage and Datasets

- 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.



Organism	*Blast Protein Count**	*Blast Peptide Count**
E.Coli	3,970	41,760
Yeast	5,684	54,948
Rat	11,962	122,322
Mouse	20,059	214,905
Human	20,843	243,749

YPED User Interface



YPED

PI Functions User: Angus Nairn

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

Project Name	Samples	Analysis Types	Organisms	PI	Date Created
16 CEX fractions Human Brain on 5600TT	17	LCMS	Homo sapiens	Angus Nairn	2012-07-31 10:14:34.0
5600TT PP2A Methyl-leucine searches, 082211	10	LCMS	Homo sapiens, Mus musculus	Angus Nairn	2011-08-22 09:21:20.0
5600TT Rat PSD - 90 min runs	1	LCMS	Rattus norvegicus	Angus Nairn	2012-01-31 13:29:28.0
5600TT methylation on PP2A	3	LCMS	Bos taurus	Angus Nairn	2012-04-24 10:46:41.0
ACID EX	8	LCMS, LabelFreeQuant	Rattus norvegicus	Angus Nairn	2012-10-26 17:30:52.0
B6 and C8 sub-cellular fractionation	10	LCMS	Homo sapiens	Angus Nairn	2012-07-31 10:04:24.0
CONT, OKA, CsA En/FT on 5600TT	6	LCMS	Homo sapiens	Angus Nairn	2011-06-21 11:36:53.0
Cocaine PSD Mouse iTRAQ	14	ITRAQ8plex, LCMS	Rattus norvegicus	Angus Nairn	2008-11-04 16:42:42.0
Combined SILAC for PF-Shari	14	LCMS	Homo sapiens	Angus Nairn	2011-03-31 11:38:34.0
DARPP32-PP1 intein	8	LCMS	Rattus norvegicus	Angus Nairn	2011-06-27 15:52:18.0

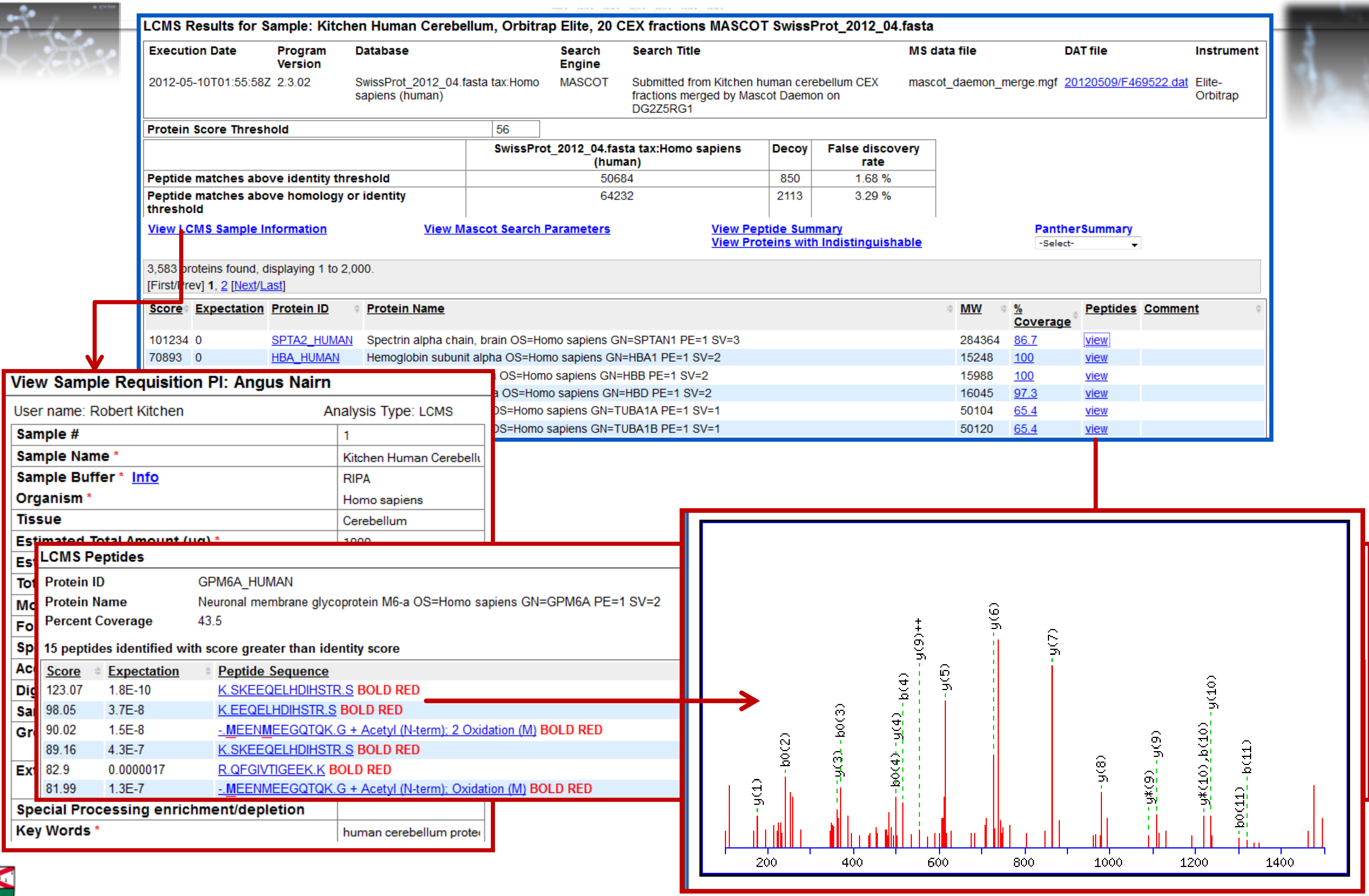
1,393 results found, displaying 181 to 190.[First/Prev] 15, 16, 17, 18, 19, 20, 21, 22 [Next/Last]

Analysis Type	Sample Name	Search Engine	Database	Date	User/PI
LCMS	Sakaue_Cortex-PSD_120611_90minIDA_1_500_3_1 info	MASCOT	1::contaminants_20090624.fasta 2::SwissProt_2012_05.fasta	2012-06-05 09:13:05.0	Fumika Sakaue/Angus Nairn
LCMS	WT#2 COC FT info	MASCOT	SwissProt_2012_05.fasta	2012-06-04 12:46:40.0	Veronica Musante/Angus Nairn
LCMS	WT#2 COC EN info	MASCOT	SwissProt_2012_05.fasta	2012-06-04 12:42:54.0	Veronica Musante/Angus Nairn
LCMS	WT#1 SAL FT info	MASCOT	SwissProt_2012_05.fasta	2012-06-04 12:35:38.0	Veronica Musante/Angus Nairn
LCMS	WT#1 SAL EN info	MASCOT	SwissProt_2012_05.fasta	2012-06-04 12:30:26.0	Veronica Musante/Angus Nairn
LCMS	HEK B5 20 CEX fractions Elite info	MASCOT	SwissProt_2012_05.fasta	2012-06-01 15:04:37.0	Robert Kitchen/Angus Nairn
LCMS	HEK C8 16 CEX fractions Elite info	MASCOT	SwissProt_2012_04.fasta	2012-06-01 09:32:14.0	Robert Kitchen/Angus Nairn
LCMS	Kitchen Human Cerebellum, Orbitrap Elite, 20 CEX fractions info	MASCOT	SwissProt_2012_04.fasta	2012-05-31 17:52:10.0	Robert Kitchen/Angus Nairn
LCMS	PSD JPT.com heavy peptides JPT database info	MASCOT	Nairn_JPT.fasta	2012-05-10 11:40:22.0	Fumika Sakaue/Angus Nairn
LCMS	PSD JPT.com heavy peptides MOUSE/RAT search info	MASCOT	SwissProt_2012_04.fasta	2012-05-03 10:59:01.0	Fumika Sakaue/Angus Nairn

Search Sample/Project Name

For questions or comments contact Mark Shifman
Updated 22 Apr 2013 16:44

YPED User Interface (Cont'd)



LCMS Results for Sample: Kitchen Human Cerebellum, Orbitrap Elite, 20 CEX fractions MASCOT SwissProt_2012_04.fasta

Execution Date	Program Version	Database	Search Engine	Search Title	MS data file	DAT file	Instrument
2012-05-10T01:55:58Z	2.3.02	SwissProt_2012_04.fasta tax:Homo sapiens (human)	MASCOT	Submitted from Kitchen human cerebellum CEX fractions merged by Mascot Daemon on DG2Z5RG1	mascot_daemon_merge.mgf	20120509/F469522.dat	Elite-Orbitrap

Protein Score Threshold	56

	SwissProt_2012_04.fasta tax:Homo sapiens (human)	Decoy	False discovery rate
Peptide matches above identity threshold	50684	850	1.68 %
Peptide matches above homology or identity threshold	64232	2113	3.29 %

View LCMS Sample Information View Mascot Search Parameters View Peptide Summary View Proteins with Indistinguishable PantherSummary

-Select-

3,583 proteins found, displaying 1 to 2,000.
[First/Prev] **1**, 2 [Next/Last]

Score	Expectation	Protein ID	Protein Name	MW	% Coverage	Peptides	Comment
101234	0	SPTA2_HUMAN	Spectrin alpha chain, brain OS=Homo sapiens GN=SPTAN1 PE=1 SV=3	284364	86.7	view	
70893	0	HBA_HUMAN	Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2	15248	100	view	
			OS=Homo sapiens GN=HBB PE=1 SV=2	15988	100	view	
			a OS=Homo sapiens GN=HBD PE=1 SV=2	16045	97.3	view	
			OS=Homo sapiens GN=TUBA1A PE=1 SV=1	50104	65.4	view	
			OS=Homo sapiens GN=TUBA1B PE=1 SV=1	50120	65.4	view	

View Sample Requisition PI: Angus Nairn

User name: Robert Kitchen Analysis Type: LCMS

Sample #	1
Sample Name *	Kitchen Human Cerebellu
Sample Buffer * Info	RIPA
Organism *	Homo sapiens
Tissue	Cerebellum
Estimated Total Amount (ug) *	1000
Es	
Tot	
Mo	
Fo	
Sp	
Ac	
Dig	
Sa	
Gr	
Ex	
Special Processing enrichment/depletion	
Key Words *	human cerebellum prote

LCMS Peptides

Protein ID	GPM6A_HUMAN
Protein Name	Neuronal membrane glycoprotein M6-a OS=Homo sapiens GN=GPM6A PE=1 SV=2
Percent Coverage	43.5

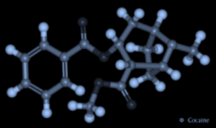
15 peptides identified with score greater than identity score

Score	Expectation	Peptide Sequence
123.07	1.8E-10	K.SKEEQELHDIHSTR.S BOLD RED
98.05	3.7E-8	K.EEQELHDIHSTR.S BOLD RED
90.02	1.5E-8	-.MEENMEEGQTQK.G + Acetyl (N-term); 2 Oxidation (M) BOLD RED
89.16	4.3E-7	K.SKEEQELHDIHSTR.S BOLD RED
82.9	0.0000017	R.QFGIVTIGEEK.K BOLD RED
81.99	1.3E-7	-.MEENMEEGQTQK.G + Acetyl (N-term); Oxidation (M) BOLD RED

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