Library of Integrated Network-based Cellular Signatures

Proteomic Characterization Center for Signaling and Epigenetics

Scientific Advisory Board Meeting

August 31, 2016

Jacob D. Jaffe – Broad Institute, PI

Li-Huei Tsai – MIT, co-investigator

Michael MacCoss – U. of Washington, co-investigator

Agenda

- Introductions
- SAB Expectations and Guidance
- Center Overview – Jake Jaffe
- Neurobiology Focus – Li-Huei Tsai
- Next Generation MS – Mike MacCoss

Q & A is welcome at any point

LINCS Consortium Structure

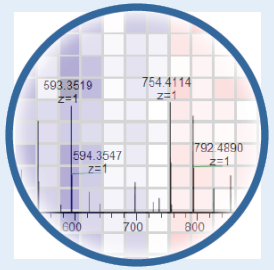
DATA GENERATION

DToxS
Mount Sinai
School of Medicine

HMS
LINCS

MEP-LINCS
OHSU

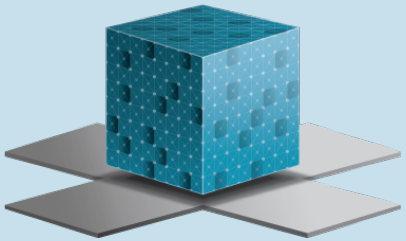
NEUROLINCS



LINCS PCCSE

LINCS Center for
Transcriptomics

DATA COORDINATION



BD2K-LINCS
DATA COORDINATION AND
INTEGRATION CENTER

PROGRAM OVERSIGHT

NIH National Institutes of Health
Turning Discovery Into Health

The role of the PCCSE Scientific Advisory Board

- Scientific and logistical guidance
- Spreading the word to the right communities
- Helping us continuously improve
- Assistance with NIH "Mid-course Review"
  - Site visit: December 9th, 2016

Mid-course review criteria

1. Feedback on Data Generation & Analysis
2. Feedback on Data Release and Accessibility
3. Feedback on Community Outreach
4. Feedback on major adaptations going forward
5. Other comments/advice

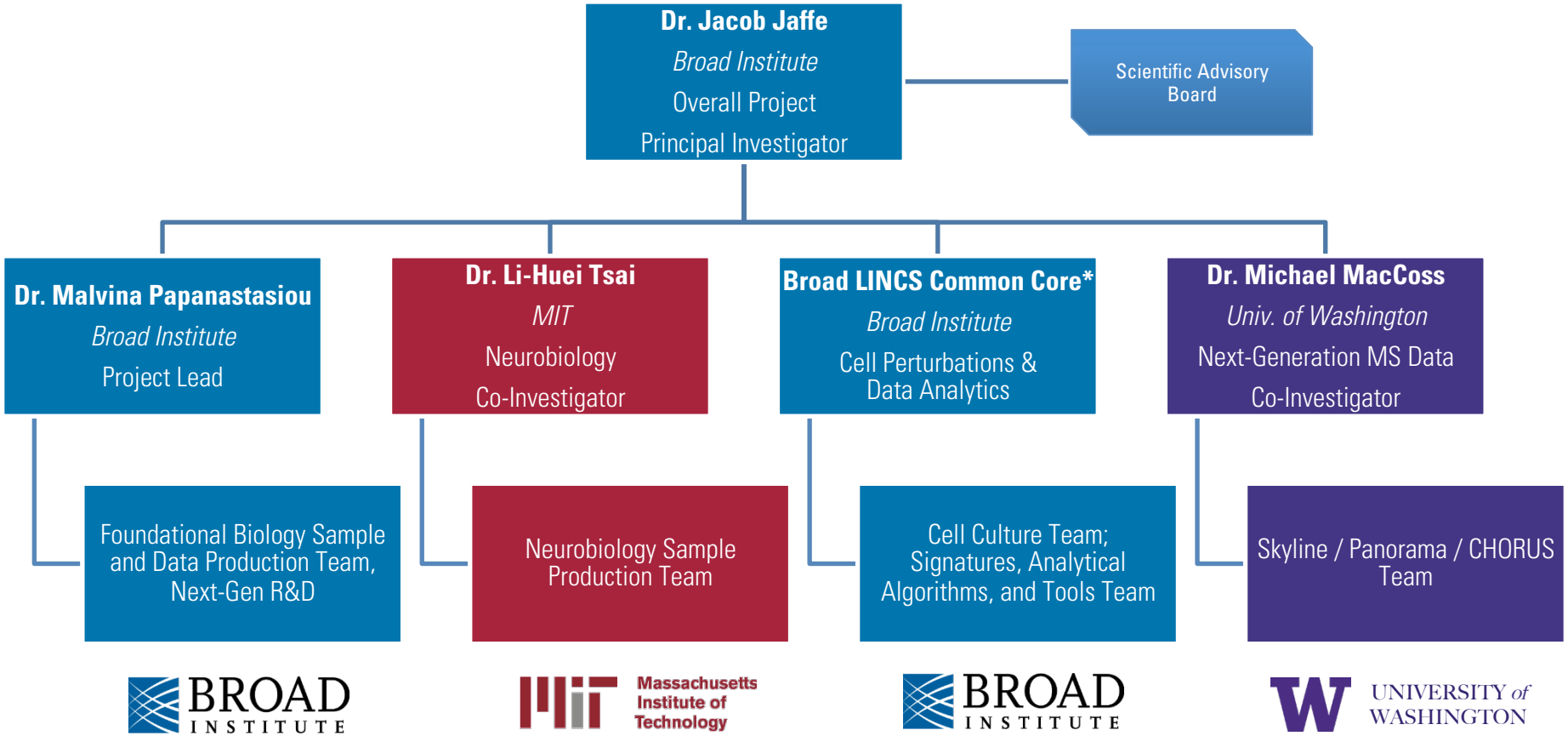
PCCSE Mission:

The LINCS Proteomic Characterization Center for Signaling and Epigenetics (PCCSE) is dedicated to understanding the therapeutic potential of drugs and unlocking the fundamental mechanisms of action of drugs, genes, and diseases through the lens of their cellular effects on phosphosignaling and epigenetic processes.

Why proteomics?

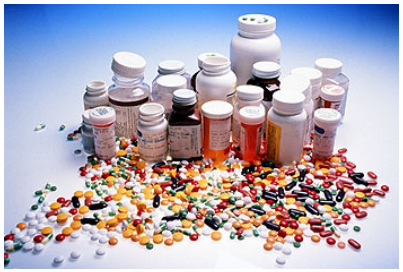
- Proteins are the targets of drugs
- Proteins mediate the actions of drugs
- Proteins are affected on time scales different than gene expression
- Proteins compose “pathways” and the bulk of molecular machinery
- Proteomics can complement other profiling methods

Organization of the Center:



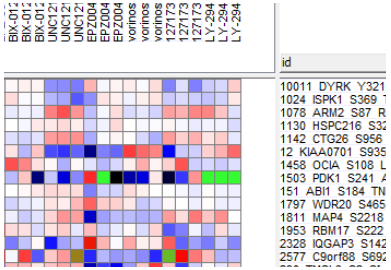
PCCSE Center Scientific Overview

Cell Lines

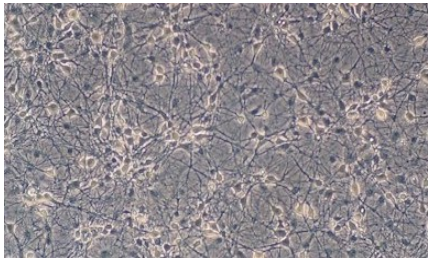


Drugs

P100 / DIA
(phospho)

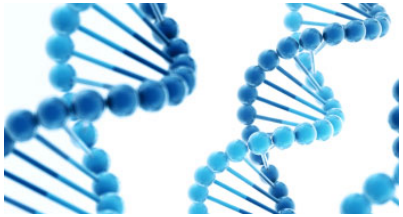


Signatures



Neural Lineages

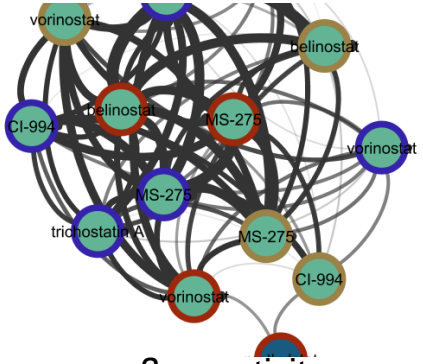
X



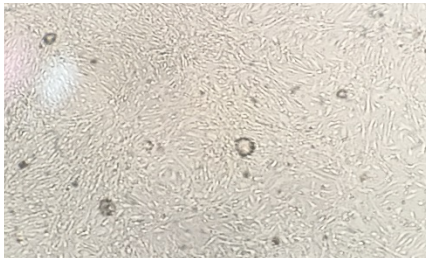
Genes

X

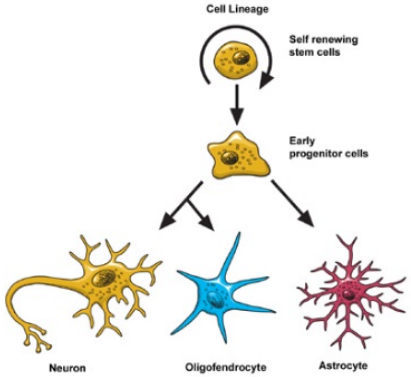
GCP
(histone marks)



Connectivity

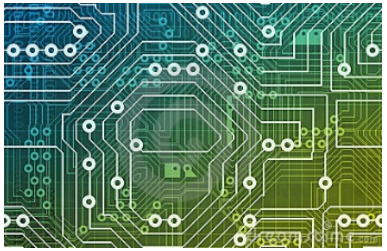


Primary Cells



Fates

L1000
(mRNA)

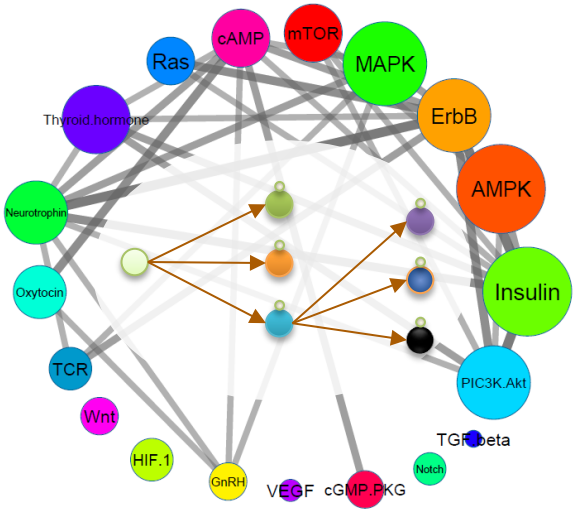


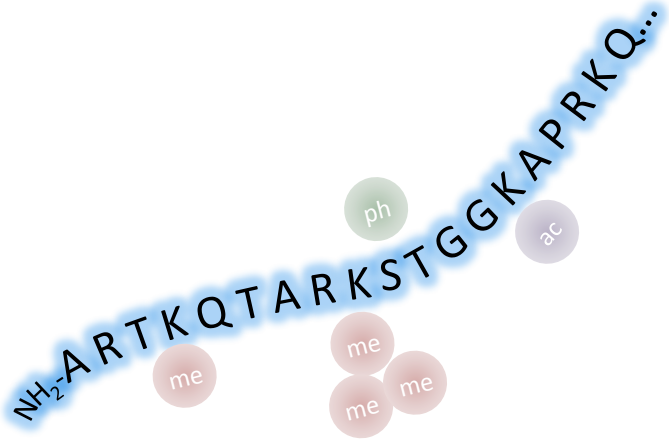
Circuitry

Proteomic Signature Assays

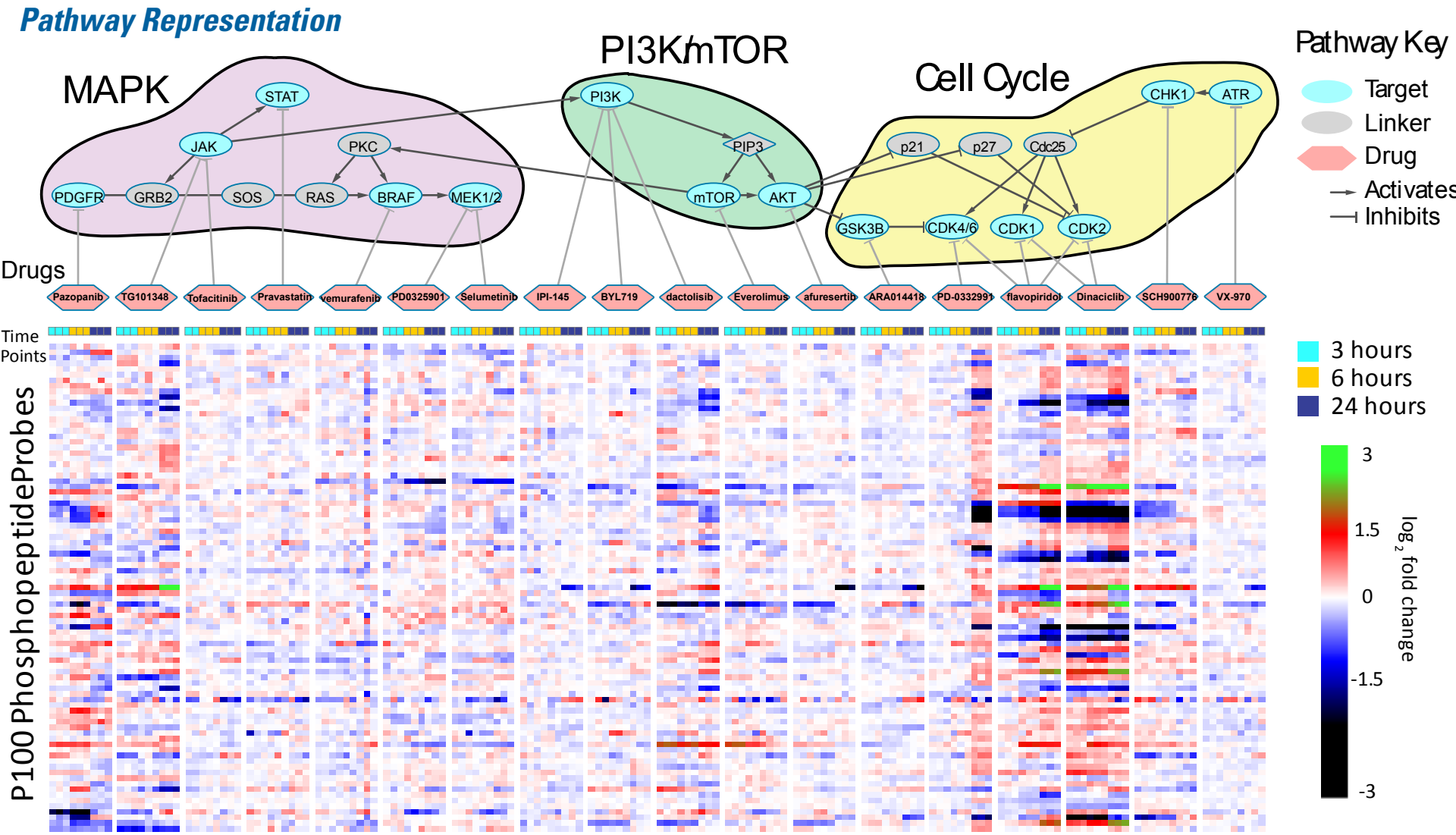
P100 is a targeted proteomic assay against 96 phosphopeptide probes that are commonly observed and modulated in diverse cell types

GCP is a targeted proteomic assay against ~ 60 probes that monitor combinations of post-translational modifications on histones



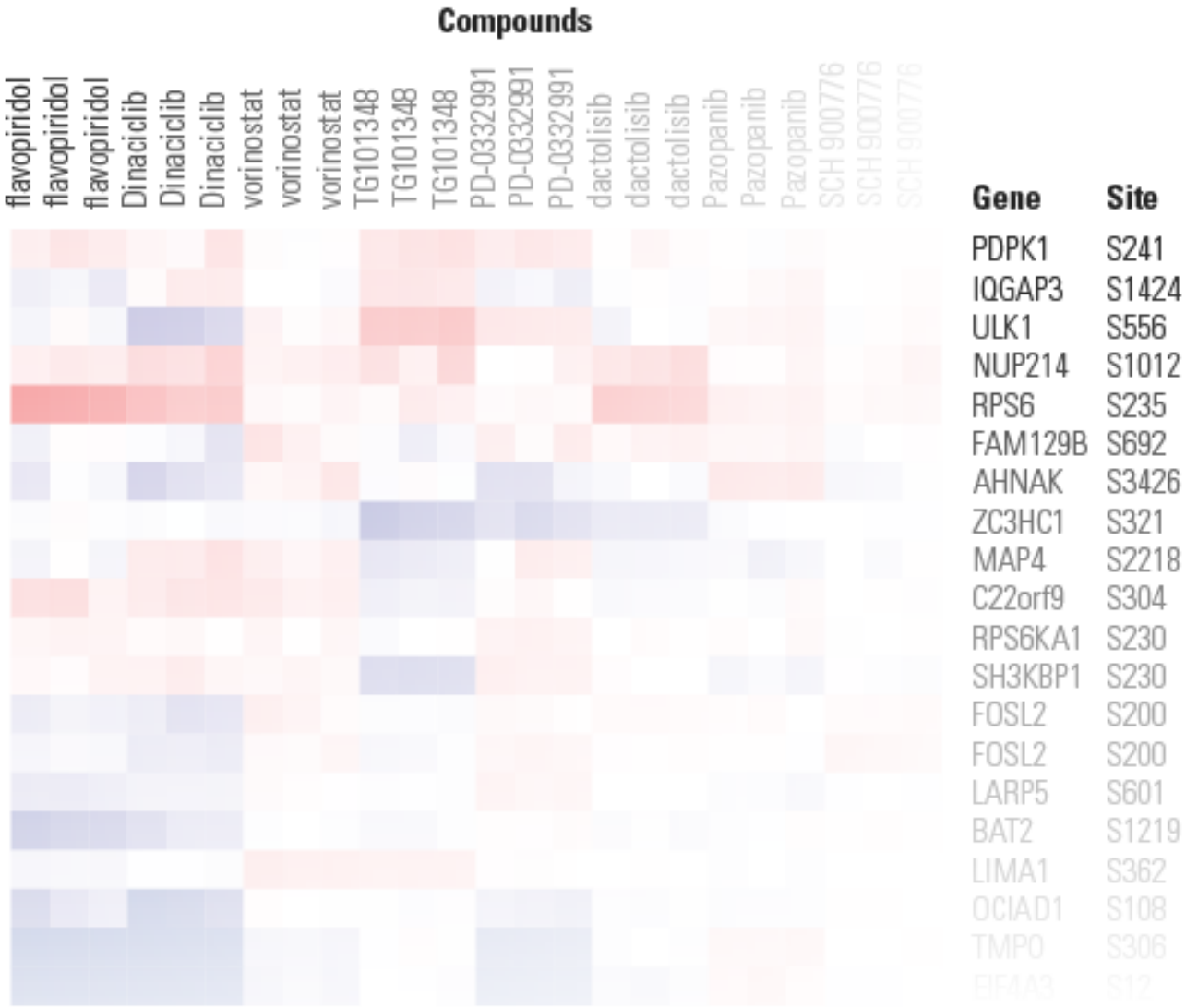


Reference signatures of drugs and pathways



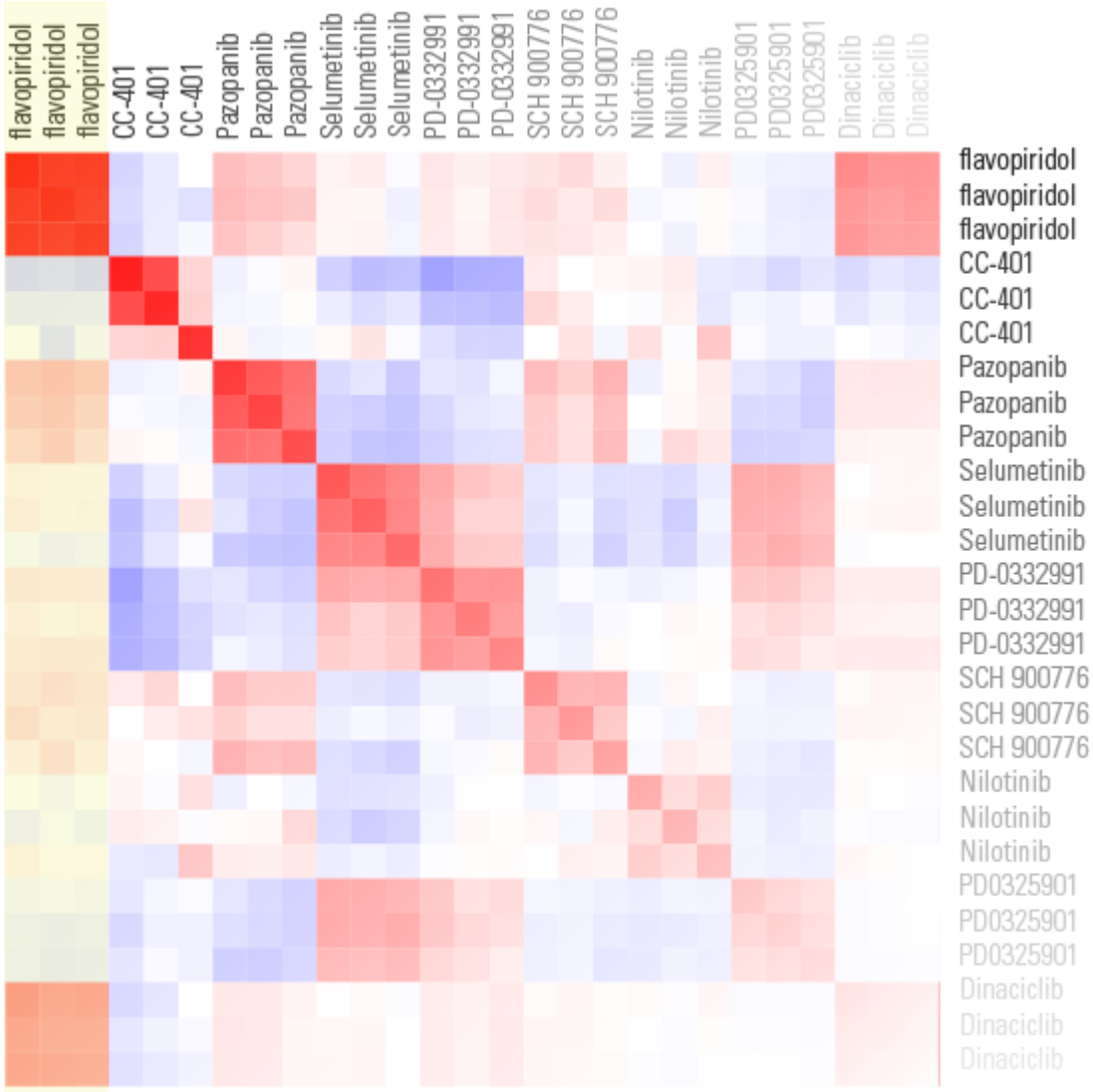
Turning signatures into connections

Signatures are groups of related profiles



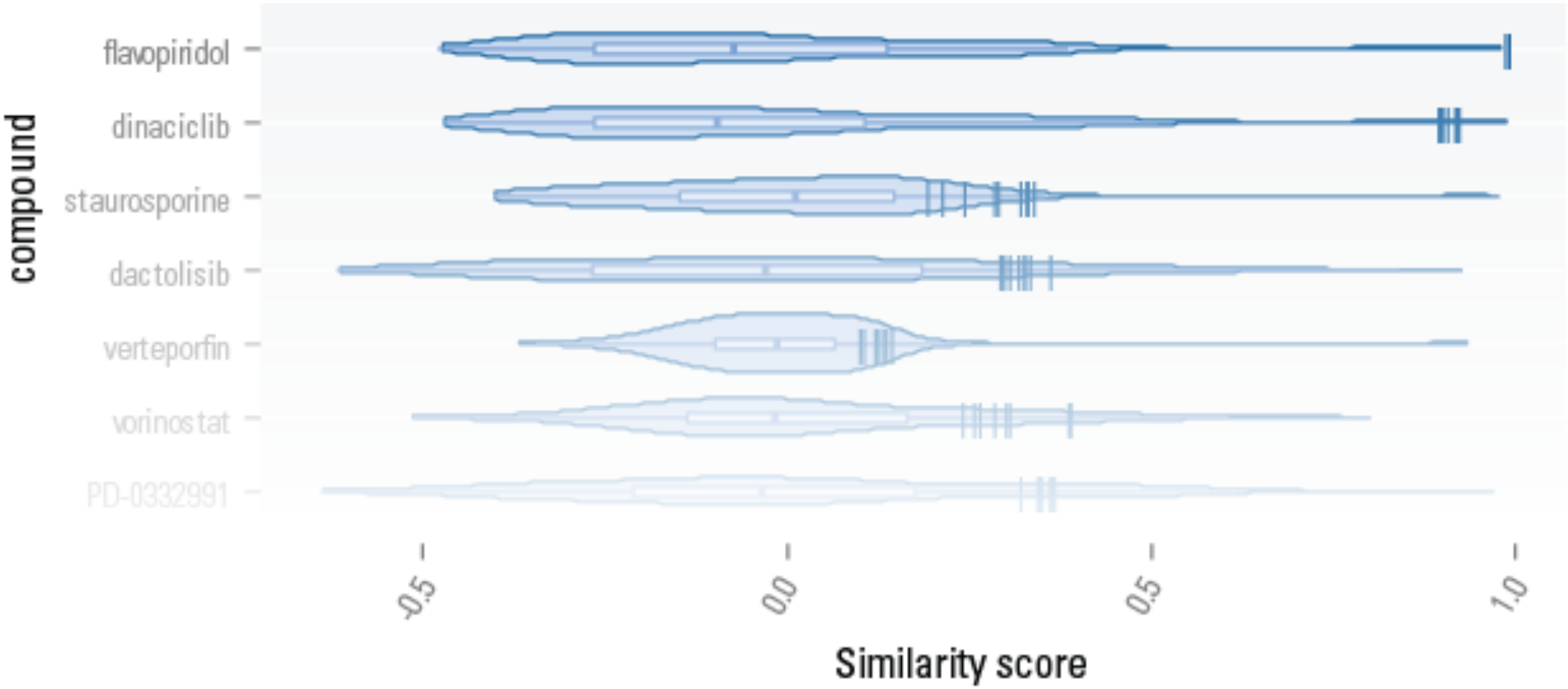
Turning signatures into connections

Similarities are correlations computed from signatures



Turning signatures into connections

Connectivity queries put correlations in context

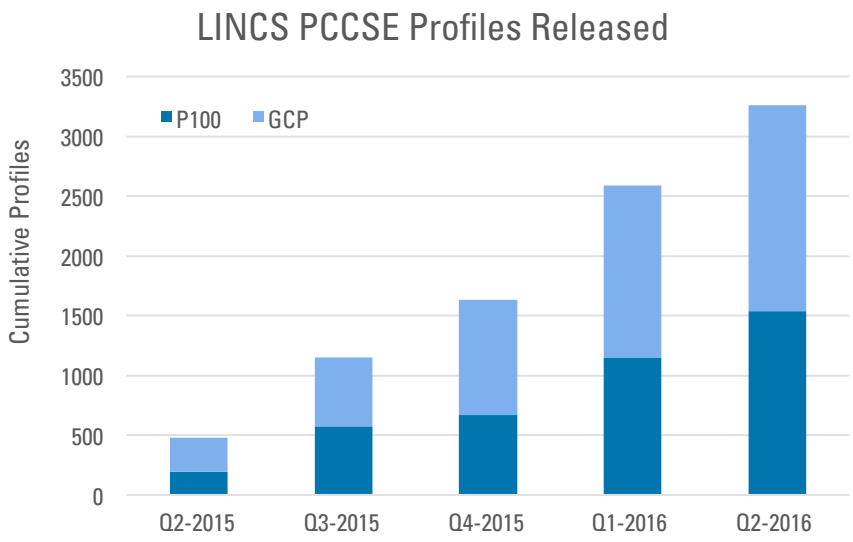


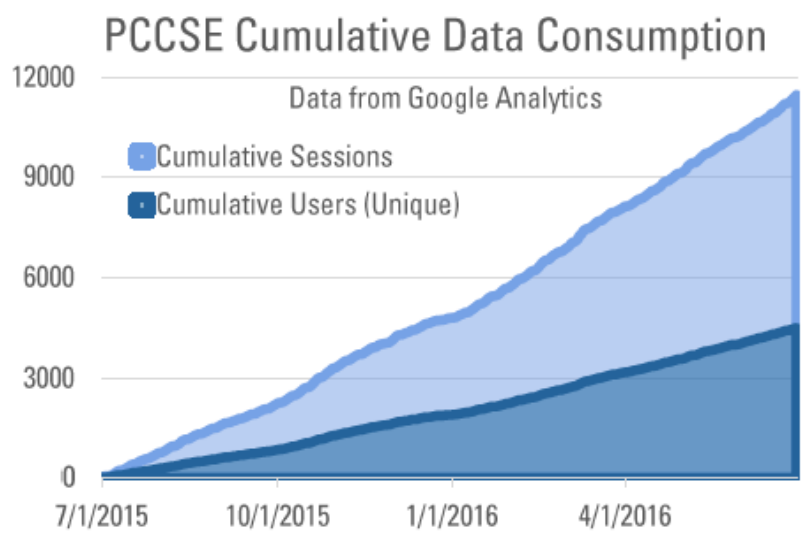
Use cases for connectivity

- What is the mechanism of action of a compound?
  - Through connectivity to drugs of known mechanism
  - Through connectivity to genetic perturbations
- Does a drug or gene's function depend on disease model or biological context?
  - Through connectivity of the same perturbation across cell types
- Does a drug activate signaling or epigenetic processes?
  - Through signature strength relative to known examples
- Are there potentially novel or unrecognized therapeutic opportunities?
  - Through *anti*-connectivity of a drug to a genetic model of disease

Extensive progress building the library

- 3400+ biological samples generated spanning:
  - 92 compounds in triplicate
  - Beginning CRISPR/Cas9-based gene disruptions
  - 6 cellular models (breast, lung, skin, prostate, and pancreatic cancers; neuronal precursors)
  - 2 assay platforms
- Key infrastructure:
  - Public data repository with built-in signature visualization (bit.ly/PCCSEData)
  - Documented workflows; automated analytical pipeline for reproducibility





Outreach activities

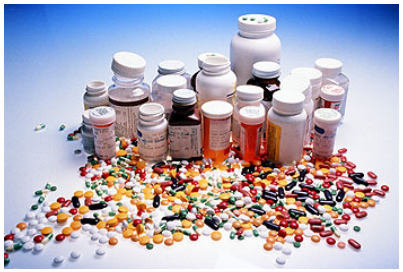
- Signatures of cardiovascular hypertension induced by chemotherapeutic drugs
  - Molecular Cardiology Research Institute, Tufts Medical Center
- Signatures of neuropsychiatric phenotypes and their responses to drugs in patient-isolated iPS-derived neurons
  - Massachusetts General Hospital
- Mapping of P100 probe-sets onto known pathway networks
  - Georgetown University and University of Delaware
  - Together with DCIC through LICNS EDSR
- Substrates to promote neuronal fates and phenotypes
  - Intra-LINCS with OHSU MEP LINCS Center

Focus areas and challenges

- Extending further into the neurobiology space
  - Li-Huei Tsai and her team will lead the way
- Unlocking Next-gen MS data
- Data: primary use, sharing, and re-use
  - What are the tools that we need to build?
  - How do we get people to utilize these data?
  - Are there audiences to be targeted and engaged?

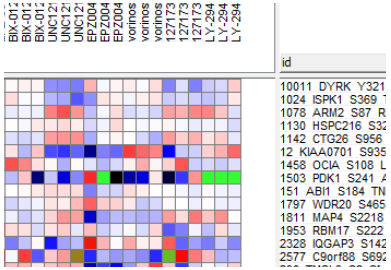
Progress and challenges in the neurobiology space

Cell Lines

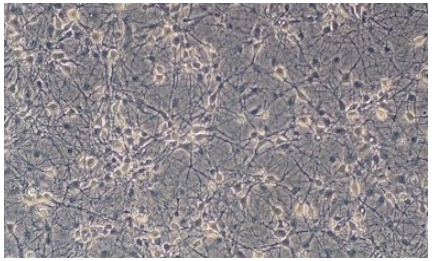


Drugs

P100 / DIA
(phospho)

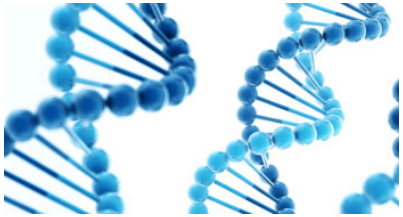


Signatures



Neural Lineages

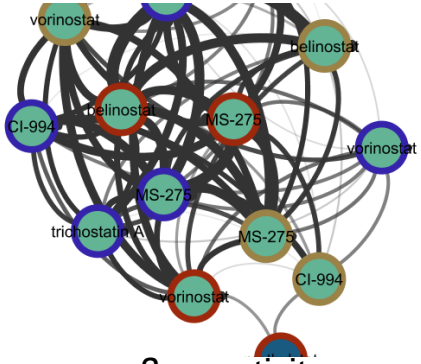
X



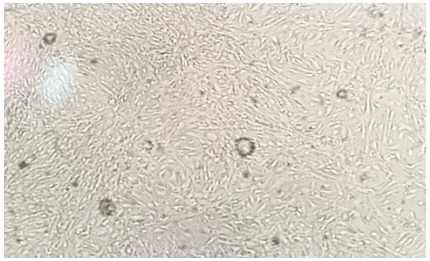
Genes

X

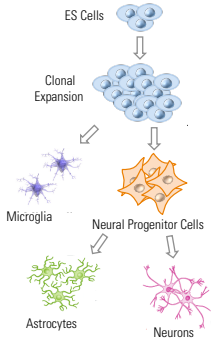
GCP
(histone marks)



Connectivity

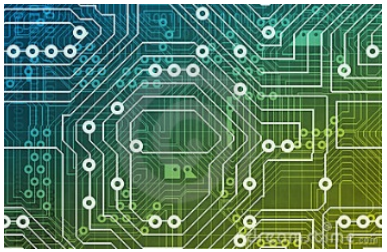


Primary Cells



Fates

L1000
(mRNA)

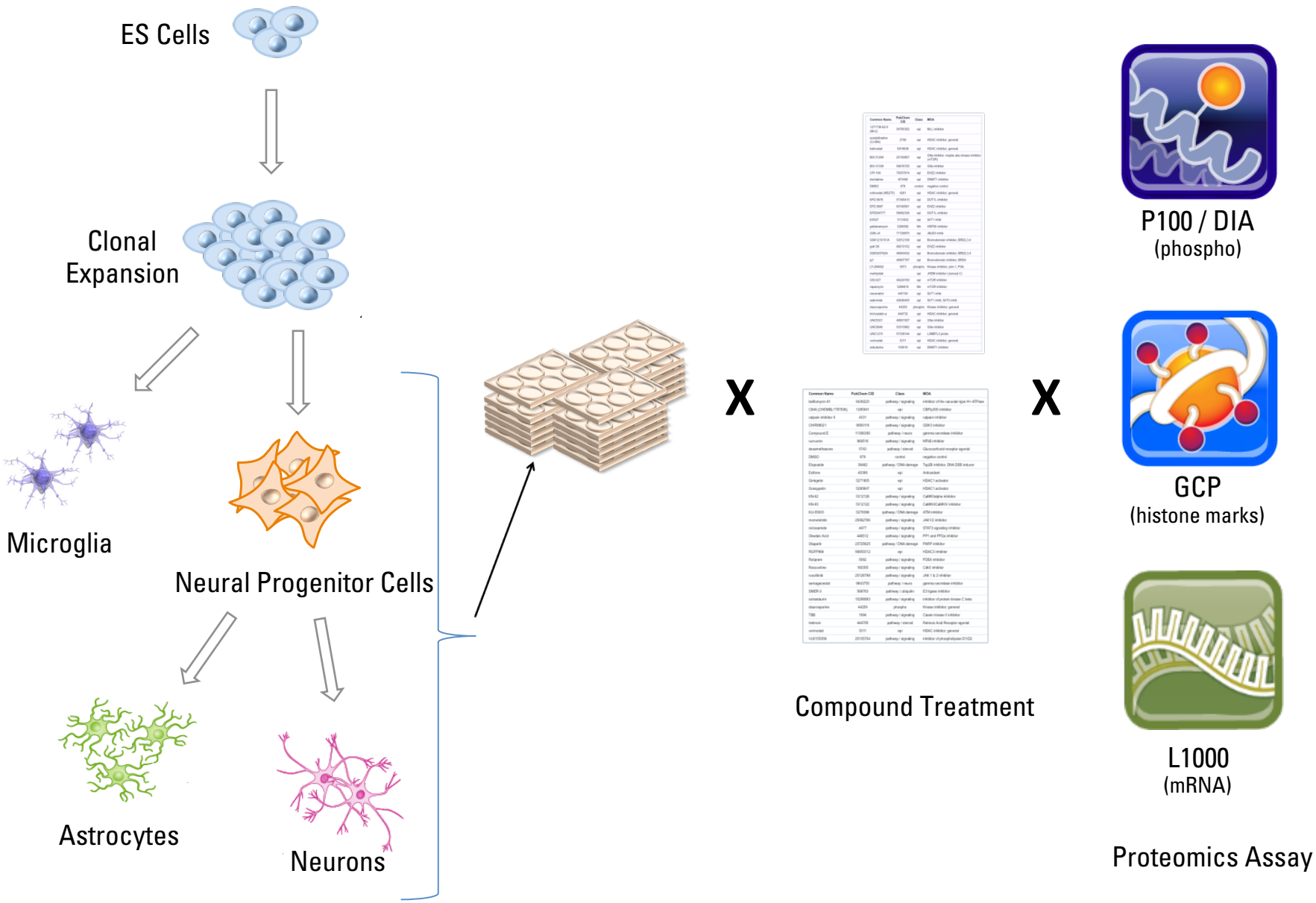


Circuitry

Neurobiology focus

- Epigenetic processes are disrupted in neurological disorders such as autism spectrum disorders (ASDs)
- Causal mutations have been found in chromatin-associated enzymes in ASD patients
- Study neural cells during development and across cellular subtypes:
  - establish epigenetic and phosphosignaling signatures
  - study effects of knocking down chromatin-modifying enzymes
  - re-create perturbations in ASD risk genes

Differences in neural cell types



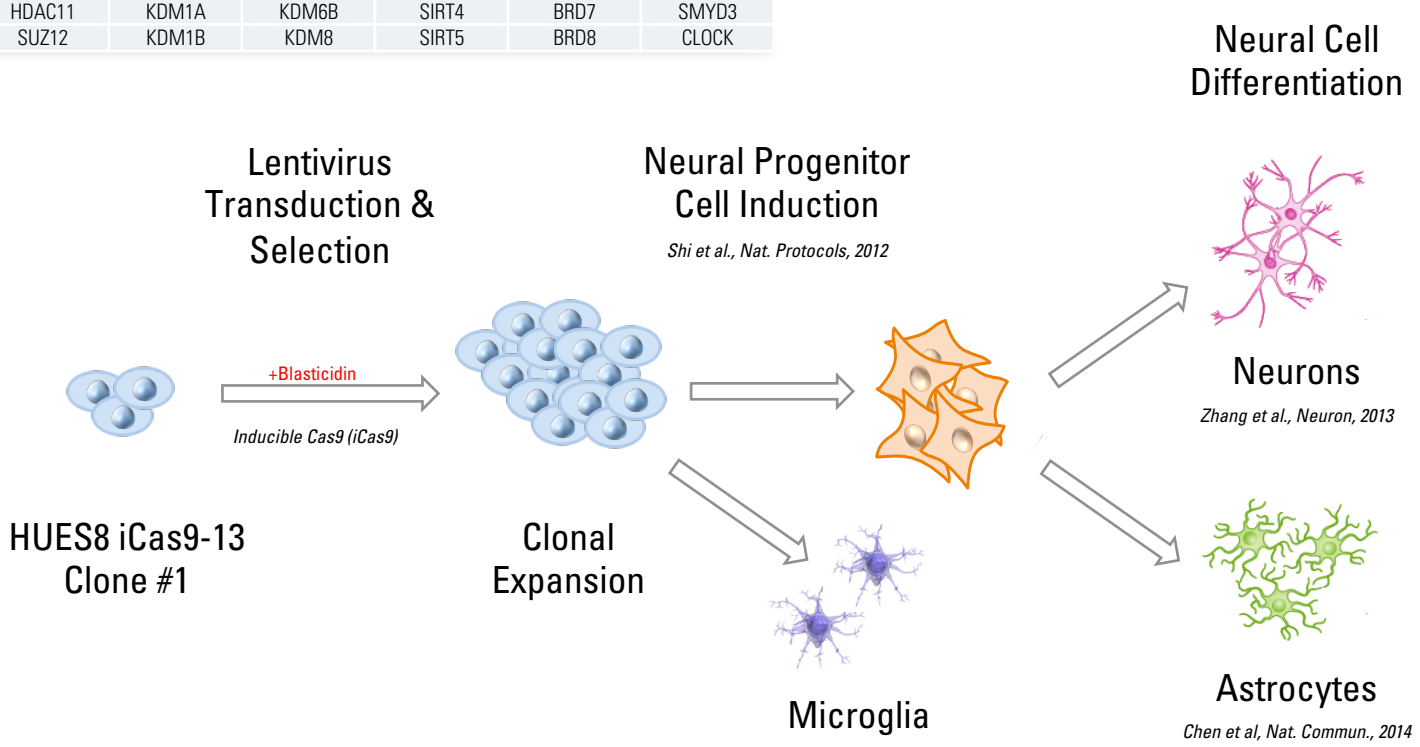
Genes to perturb neural cell types

Core Epigenetic Reference Genes

CHD8	HDAC2	KDM2A	KMT2A	SIRT6	BRD9
DNMT1	HDAC3	KDM2B	KMT2D	SIRT7	CBX1
DNMT3A	HDAC4	KDM3A	KMT2C	SUV39H1	CBX2
DNMT3AP1	HDAC5	KDM3B	KMT2E	SUV39H2	CBX3
DNMT3B	HDAC6	KDM4A	SETD1A	SUV420H1	CBX4
DNMT3L	HDAC7	KDM4B	SETD1B	SUV420H2	CBX5
DOT1L	HDAC8	KDM4C	SETD2	WHSC1	CBX6
EHMT1	KAT2A	KDM4D	SETD7	WHSC1L1	CBX7
EHMT2	KAT2B	KDM4E	SETD8	NELFA	CBX8
EZH1	KAT5	KDM5A	SETDB1	ASXL1	EED
EZH2	KAT6A	KDM5B	SETDB2	BRD1	EP300
HAT1	KAT6B	KDM5C	SIRT1	BRD2	NSD1
HDAC1	KAT7	KDM5D	SIRT2	BRD3	SMYD1
HDAC10	KAT8	KDM6A	SIRT3	BRD4	SMYD2
HDAC11	KDM1A	KDM6B	SIRT4	BRD7	SMYD3
SUZ12	KDM1B	KDM8	SIRT5	BRD8	CLOCK

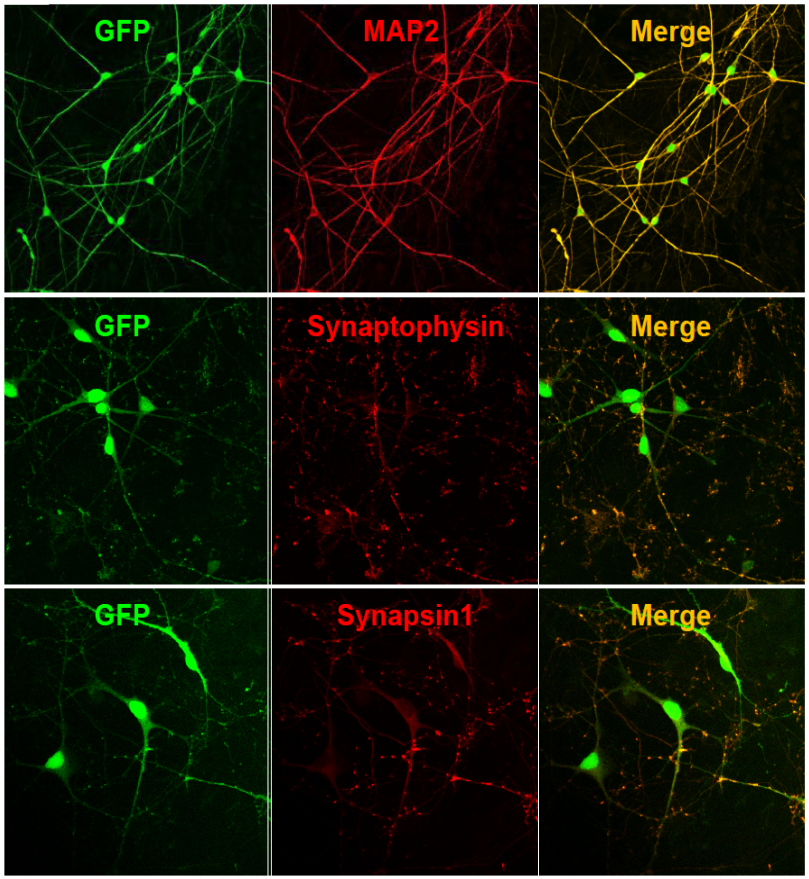
Autism-Spectrum Disorder & Other Neurobiology Genes

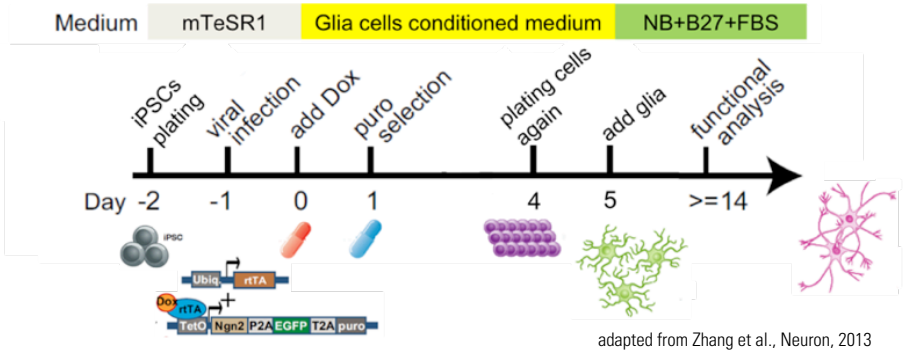
ADNP	DYRK1a	MECP2	SEMA5A
ARID1B	EHMT1	NLGN3	SHANK3
CACNA1C	EIF4G1	NLGN4X	SYNGAP1
CNTN4	EPC2	NRXN1	TLK2
CNTNAP2	FMR1	PTCHD1	TSC1
CREBBP	GRIN2B	PTEN	TSC2
CTCF	GRIP1	SCN1A	UBE3a
CTNNB1	GRM5	SCN2a	UBE3b



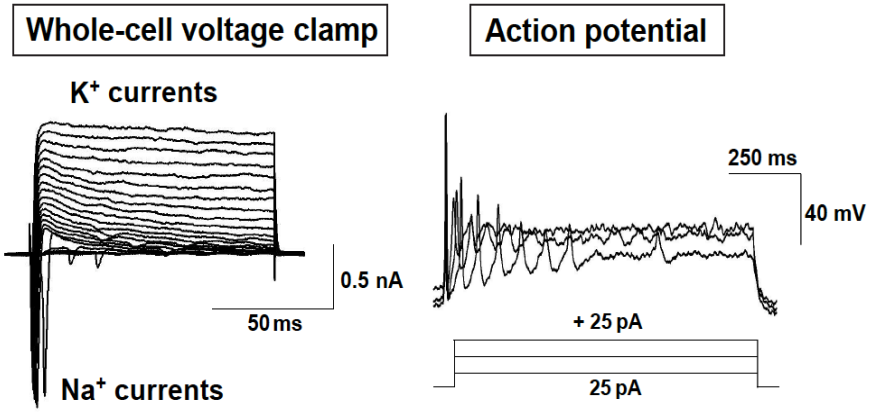
Creating neurons for high-content proteomics assays

OPTION 1: Neurogenin-2 (Ngn2) induction to force excitatory cortical neuron production



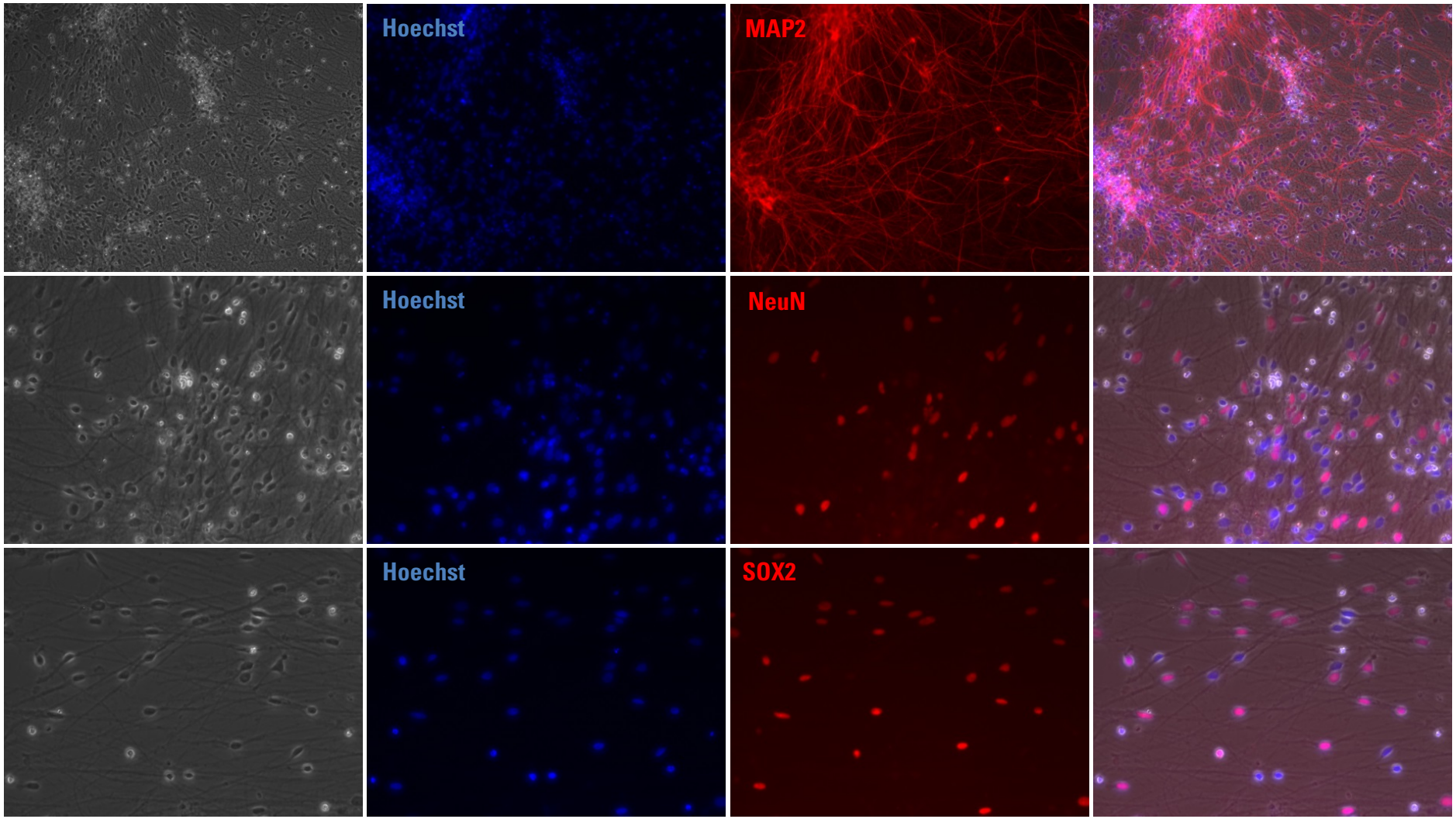


adapted from Zhang et al., Neuron, 2013



Creating neurons for high-content proteomics assays

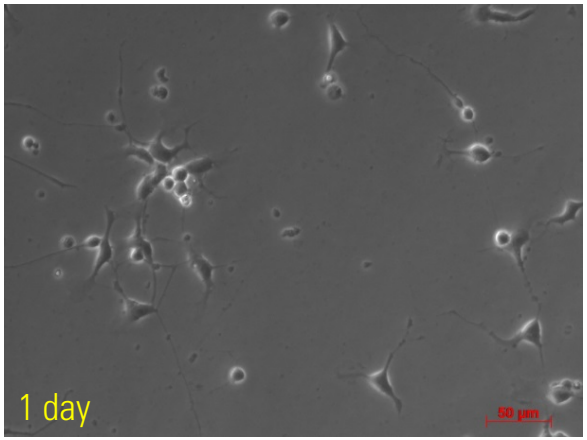
OPTION 2: Passive differentiation via growth factor withdrawal (GFW)

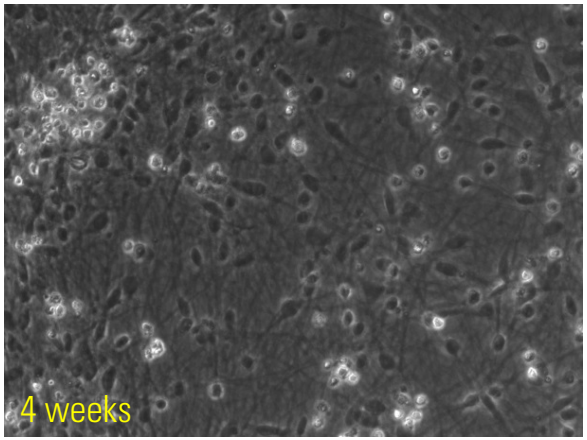


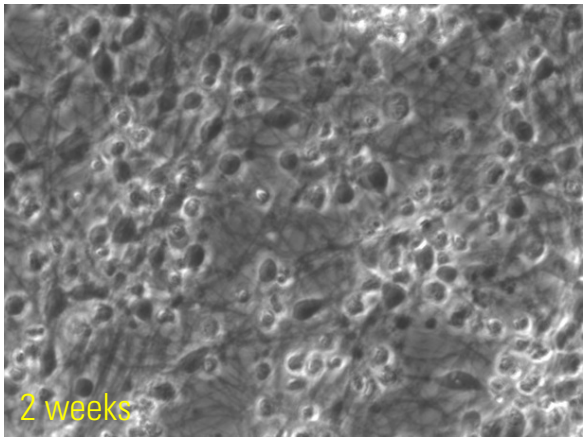
4 weeks

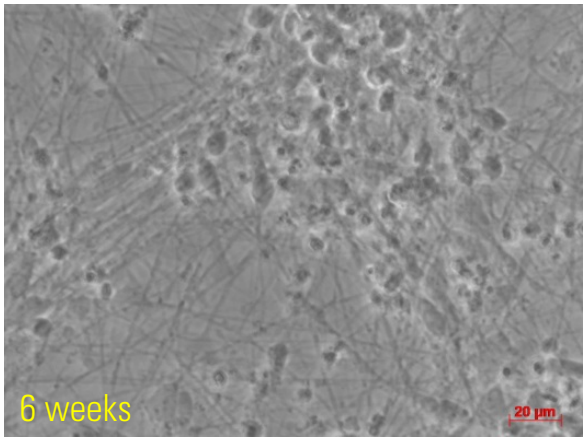
Creating neurons for high-content proteomics assays

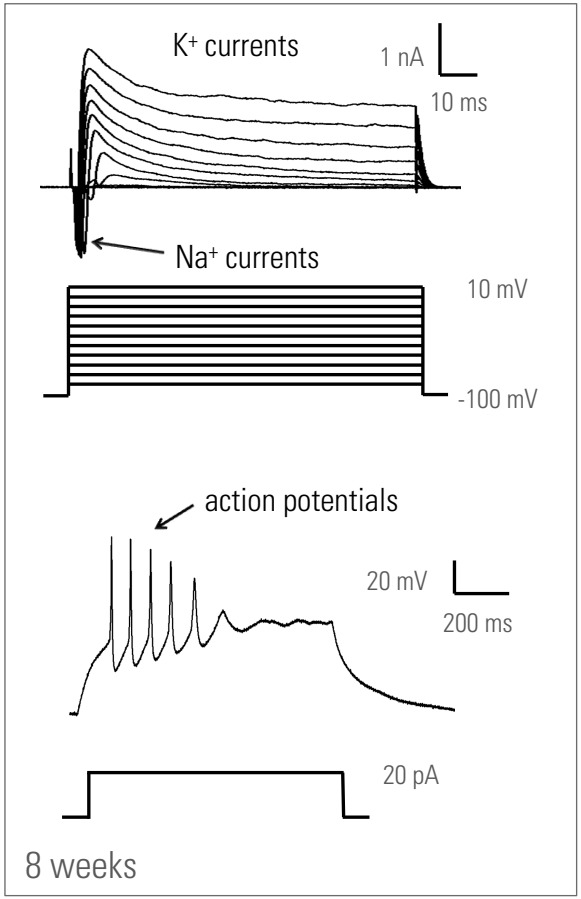
OPTION 2: Passive differentiation via growth factor withdrawal (GFW)





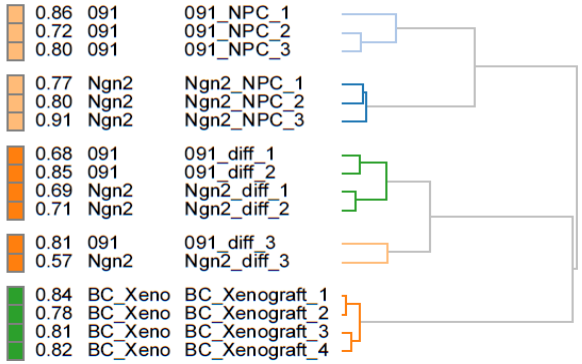


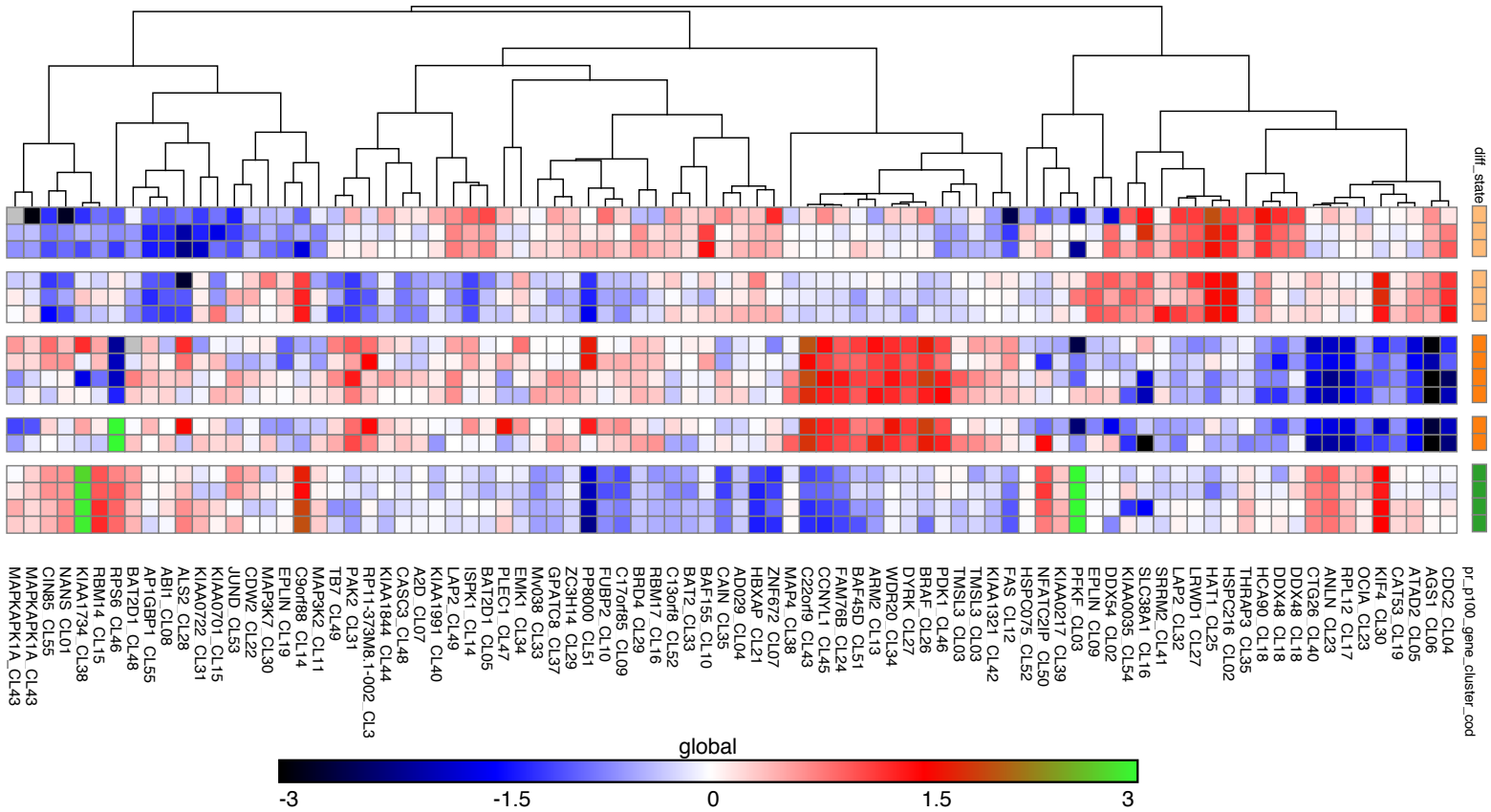




P100 comparison of different cell types

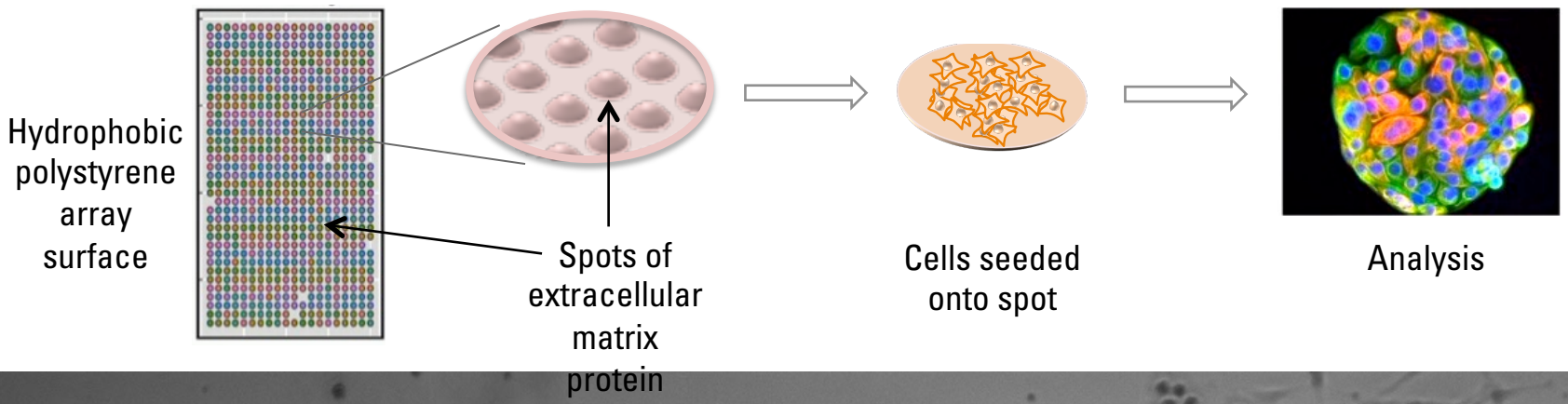
- Neural Progenitor Cells (NPCs)
- Differentiated Neurons
- Breast Cancer (BC) Xenograft line

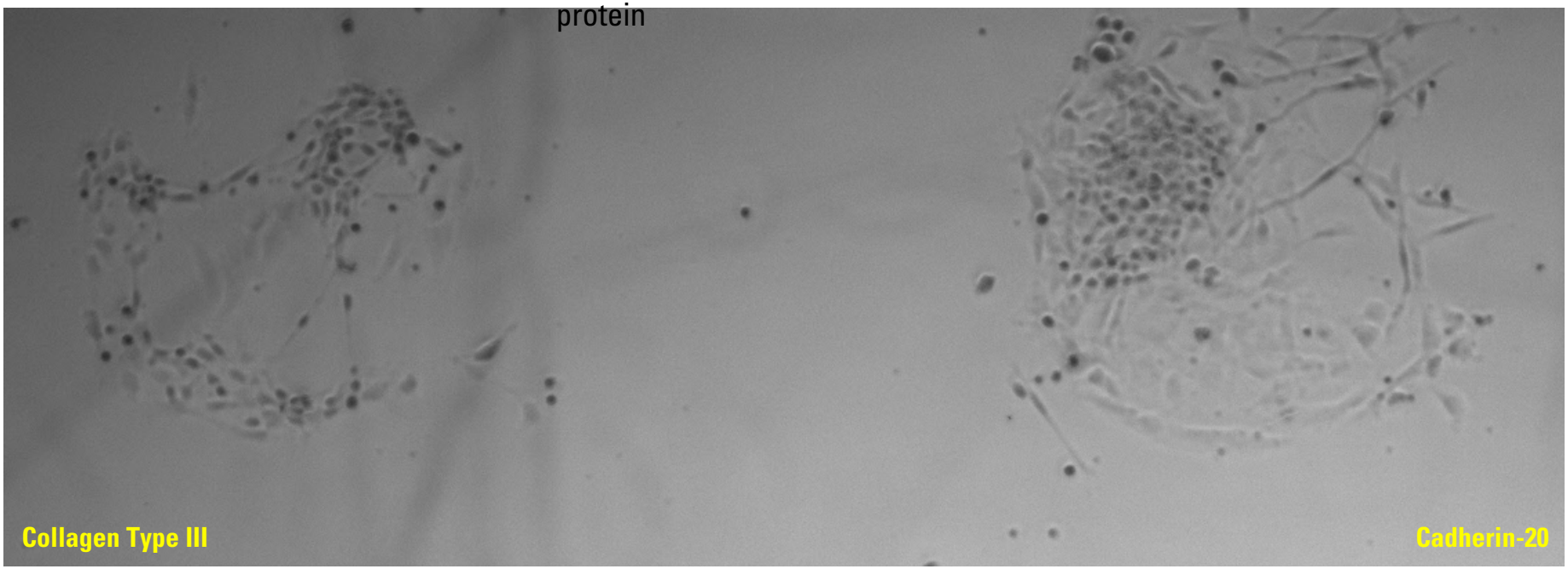




Steve Haggarty, Isaac Kohane, Roy Perlis

Microenvironment Perturbagen (MEP) LINCS Center



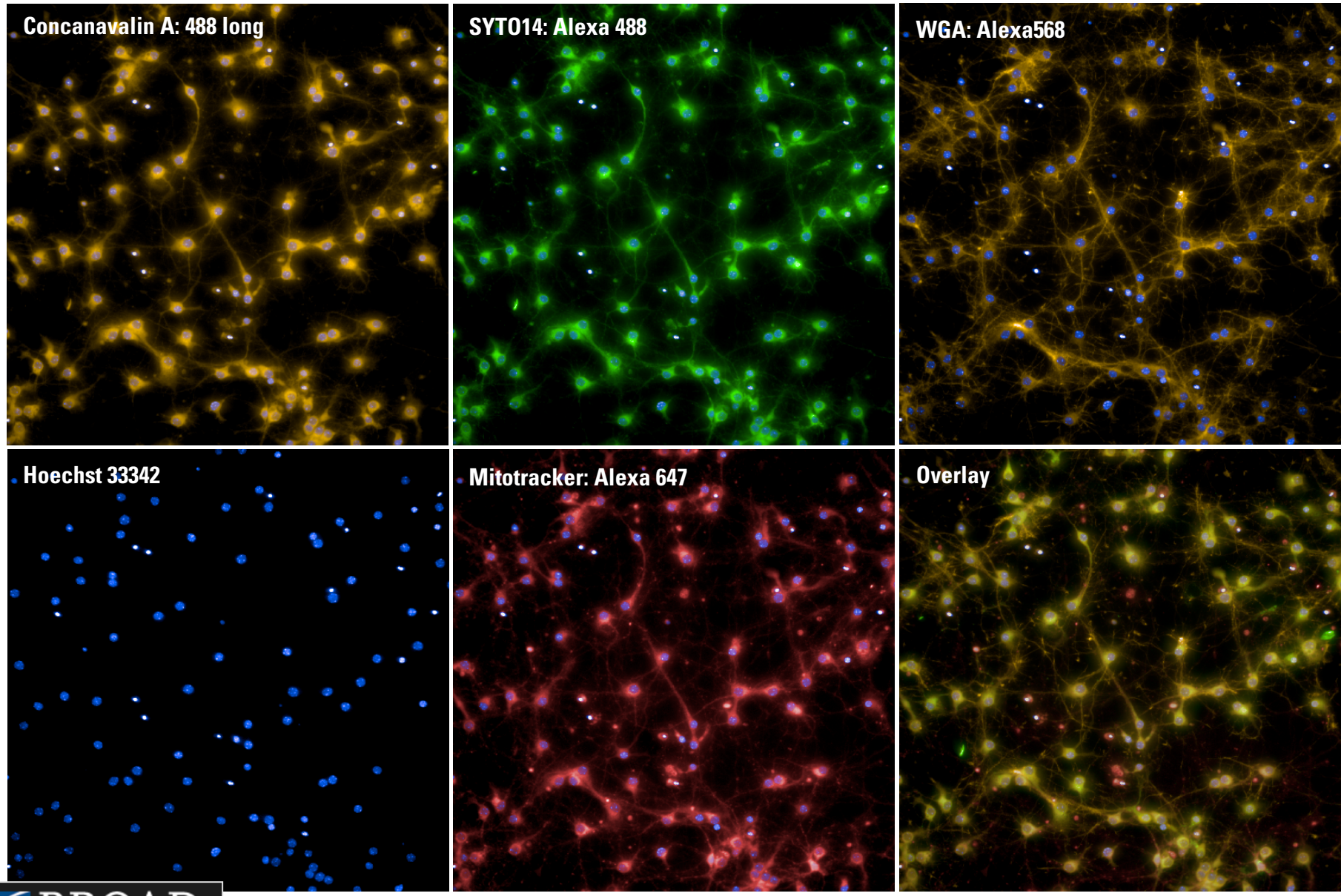


HUES8-iCAS9-13 NPCs - **TWO WEEKS** after GFW

MEP-LINCS

Joe Gray, Laura Heiser, Jim Korkola

Profiling MEMA with cell painting



BROAD INSTITUTE

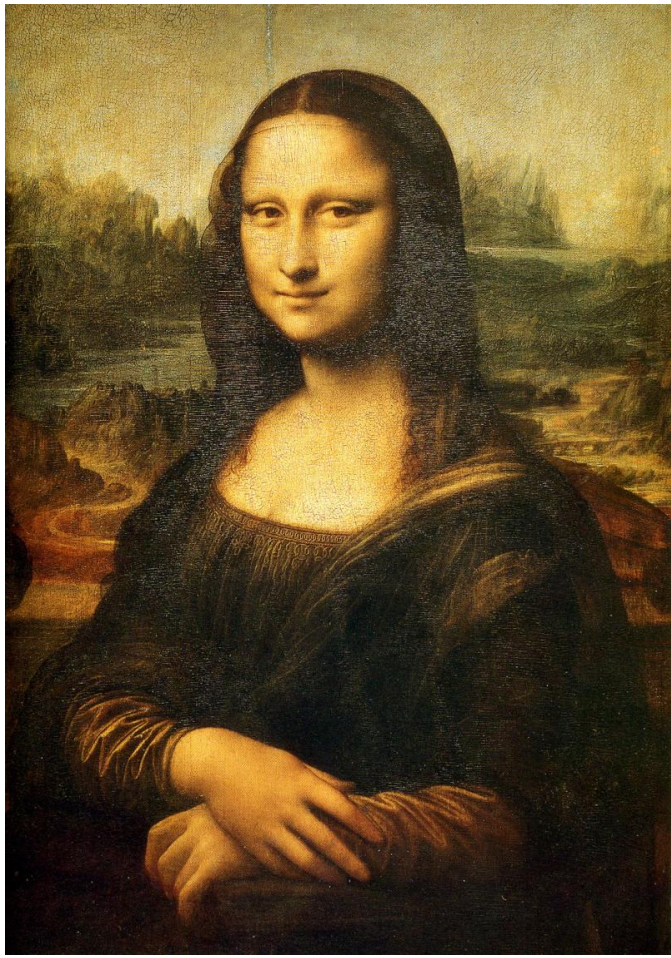
Anne Carpenter

Unlocking Next-Gen MS and promoting data reuse

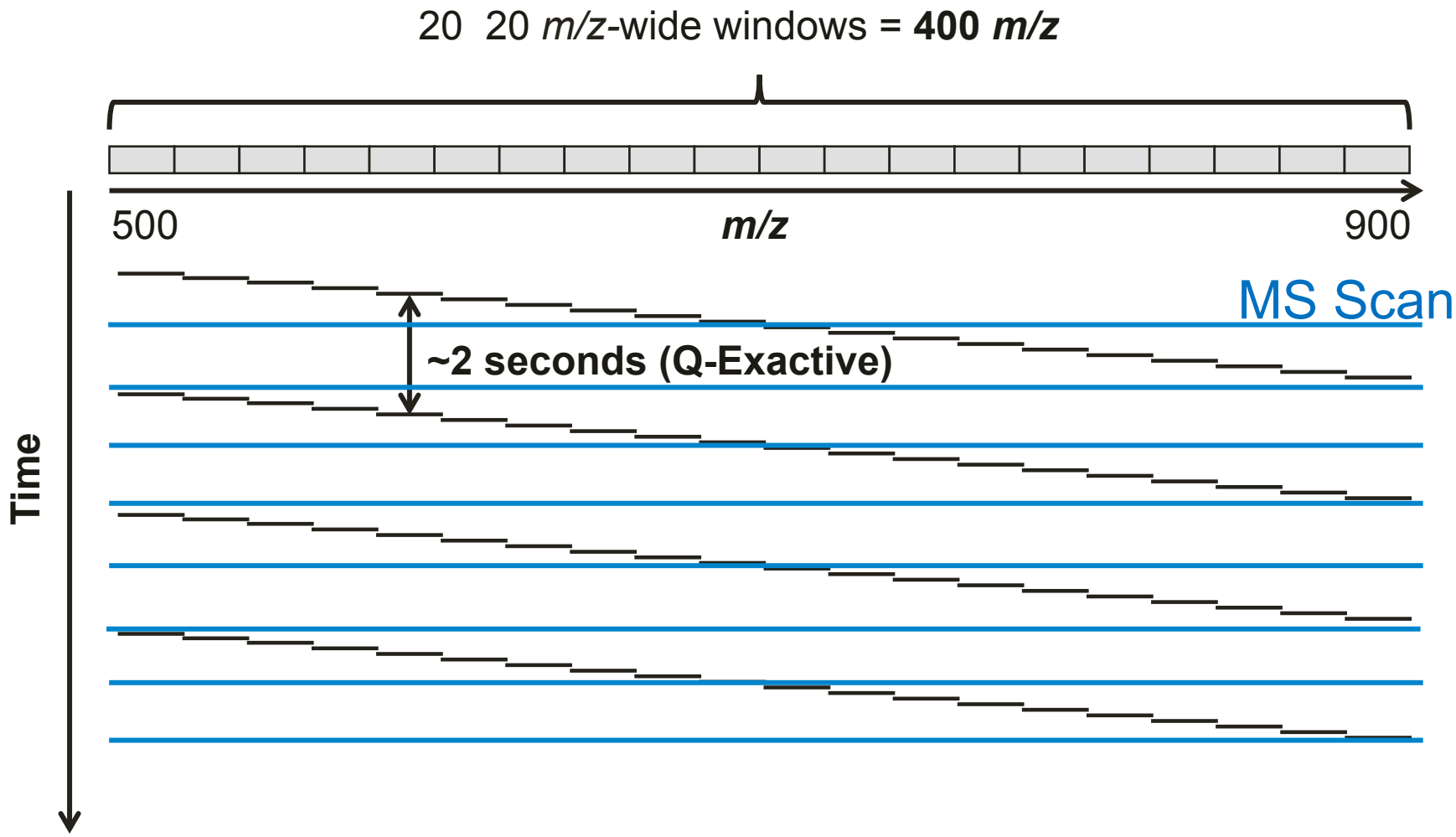
- What is Next-Gen mass spectrometry for proteomics?
- How will we unlock these data?
- How we will share our data and our methods with the world?

The Promise of Data Independent Acquisition MS (DIA)

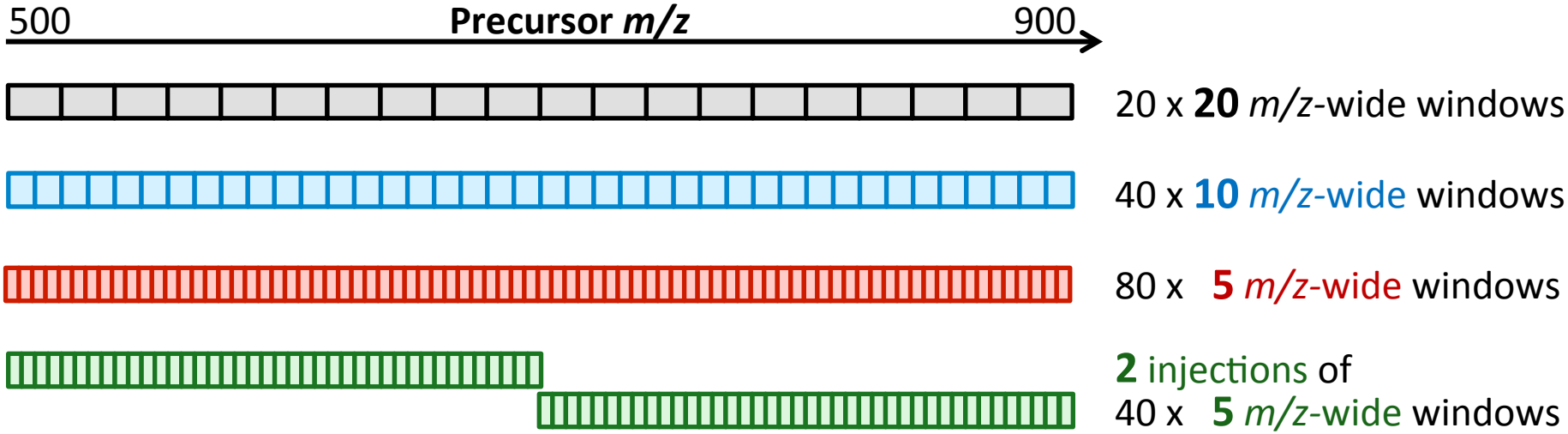
- Acquire a “molecular image” or “digital archive” of the sample
  - Mine it over and over
  - No specific targeting or scheduling
- Direct queries (and p-values) for peptides of interest
- Better quantitation than Shotgun/DDA



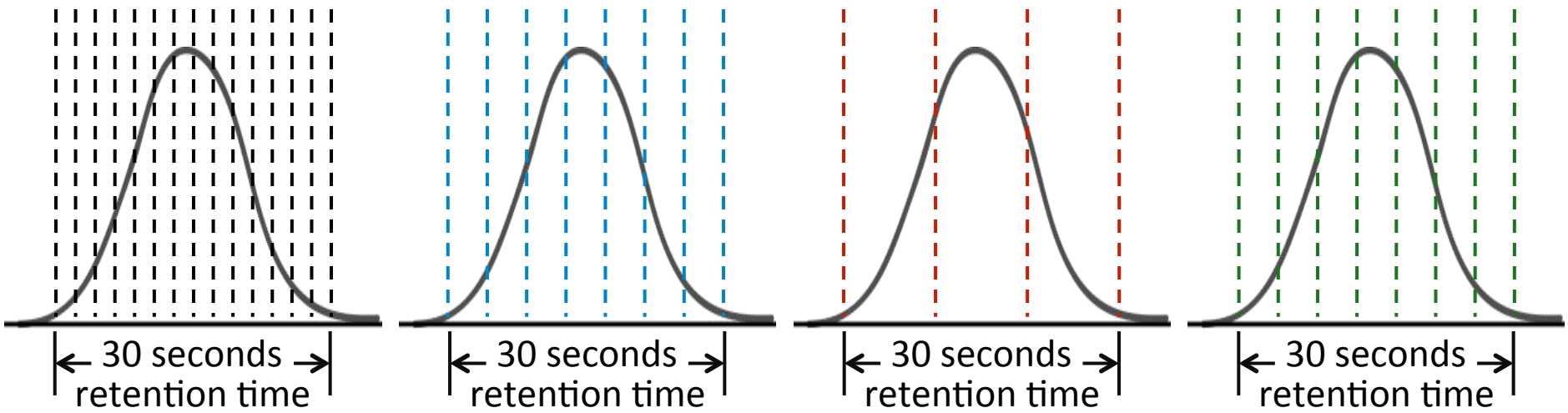
How does DIA work?



DIA is all about balance (and sacrifice) for quantification

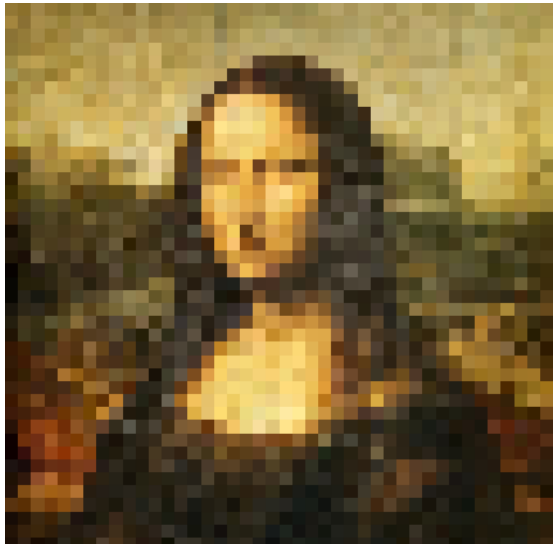


Instrument MS/MS scan speed: 10 Hz

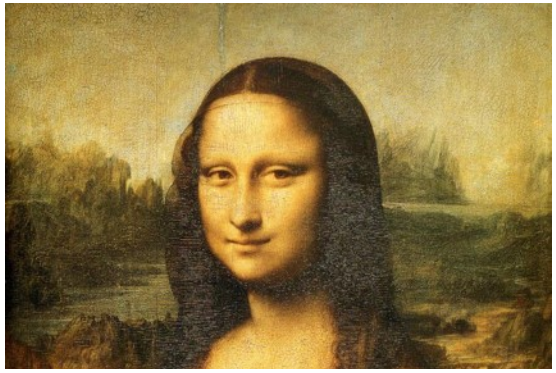


Archive Quality

20 x **20** *m/z*-wide windows

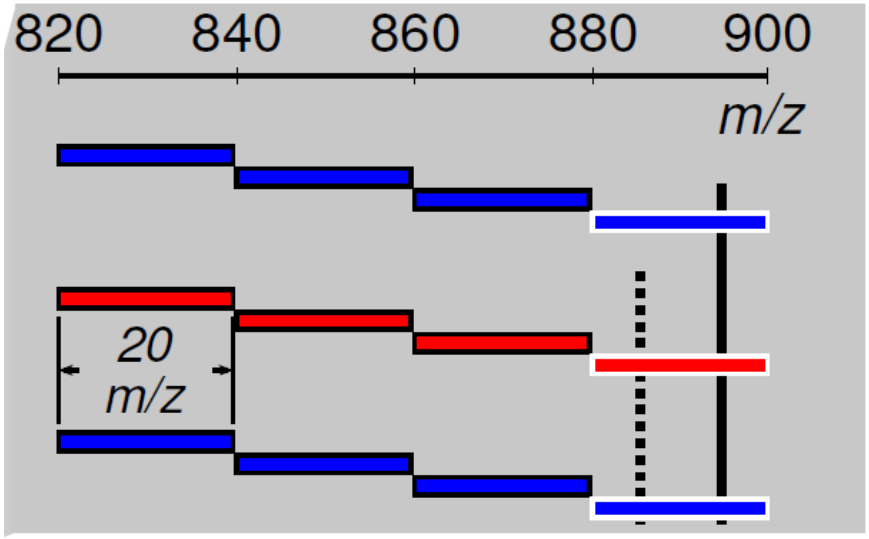


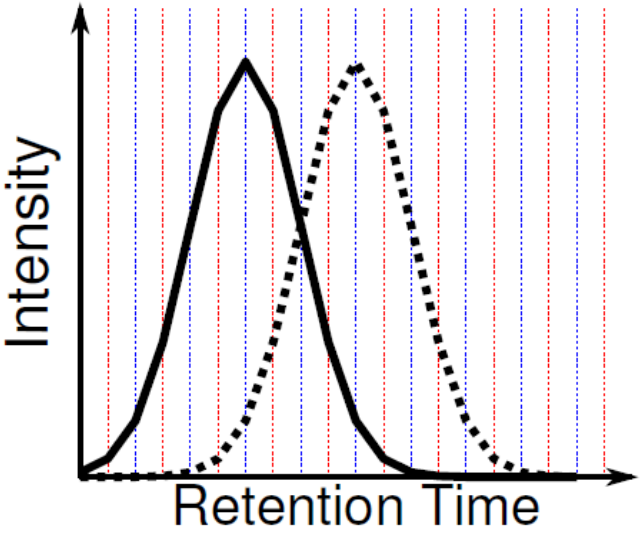
30 x **10** *m/z*-wide windows



2 injections of
40 x **5** *m/z*-wide windows

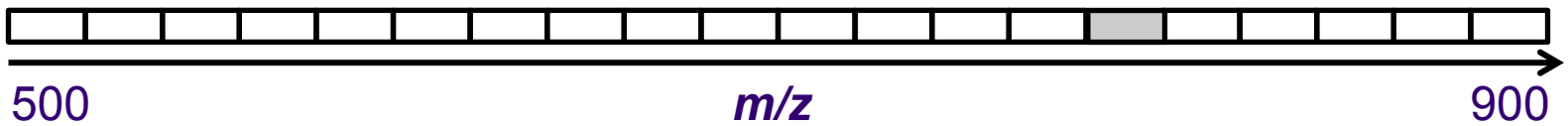
Innovative approach to improve clarity

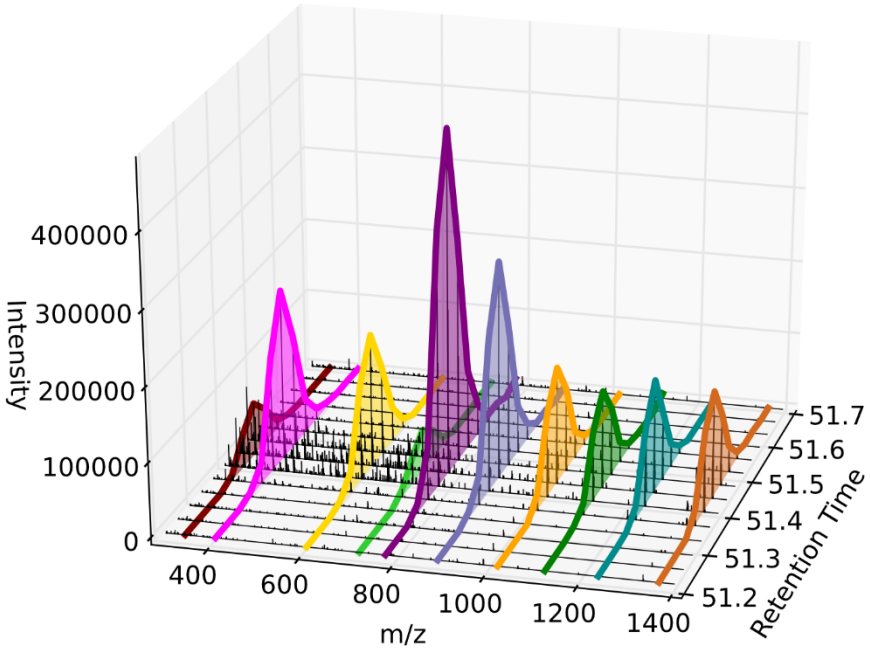


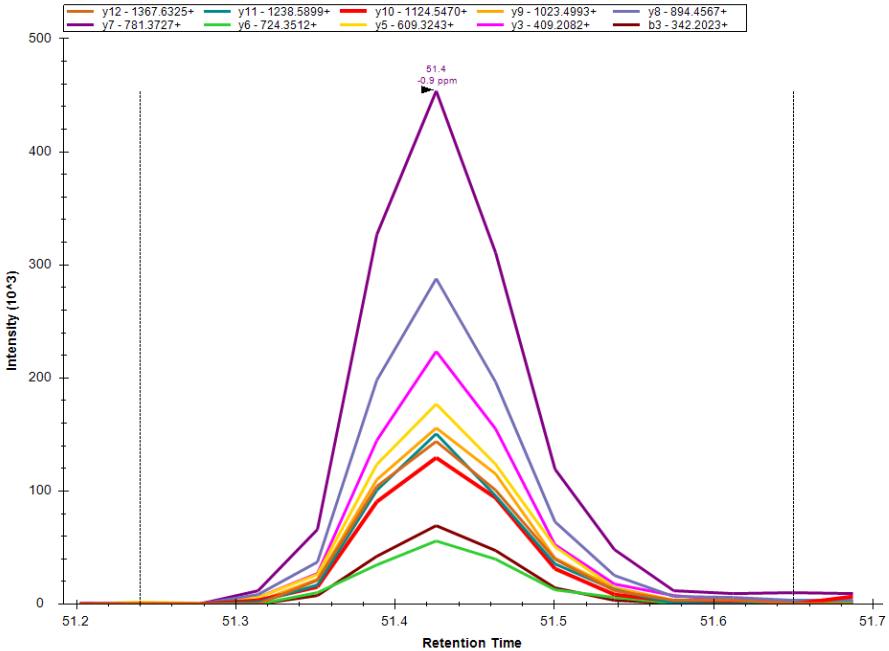


Extraction of Fragment Ions from DIA Data

VLENTFEIGSDSIFDK++ (790.4 *m/z*)

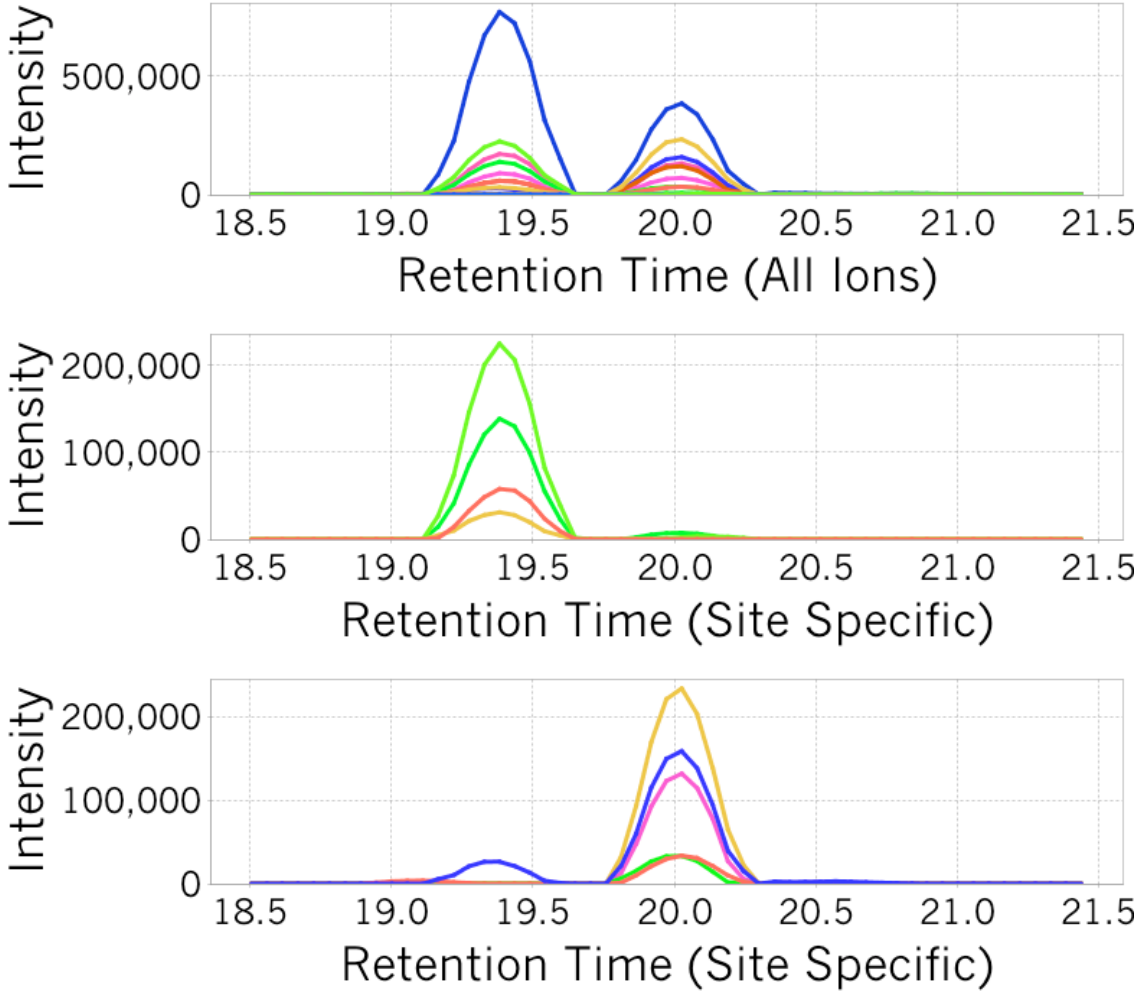




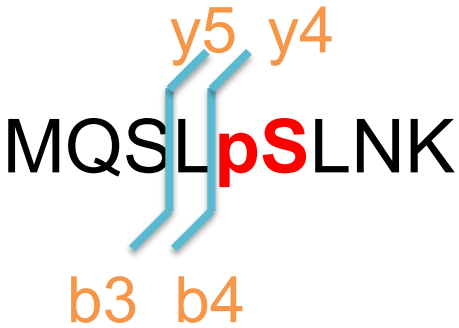


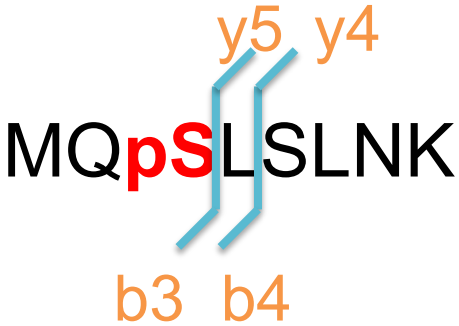
Venable et al Nat Methods 2004
Dong et al Science 2007

Localizing using site specific ions for phosphopeptides



All ions for MQSLpSLNK
or MQpSLSLNK



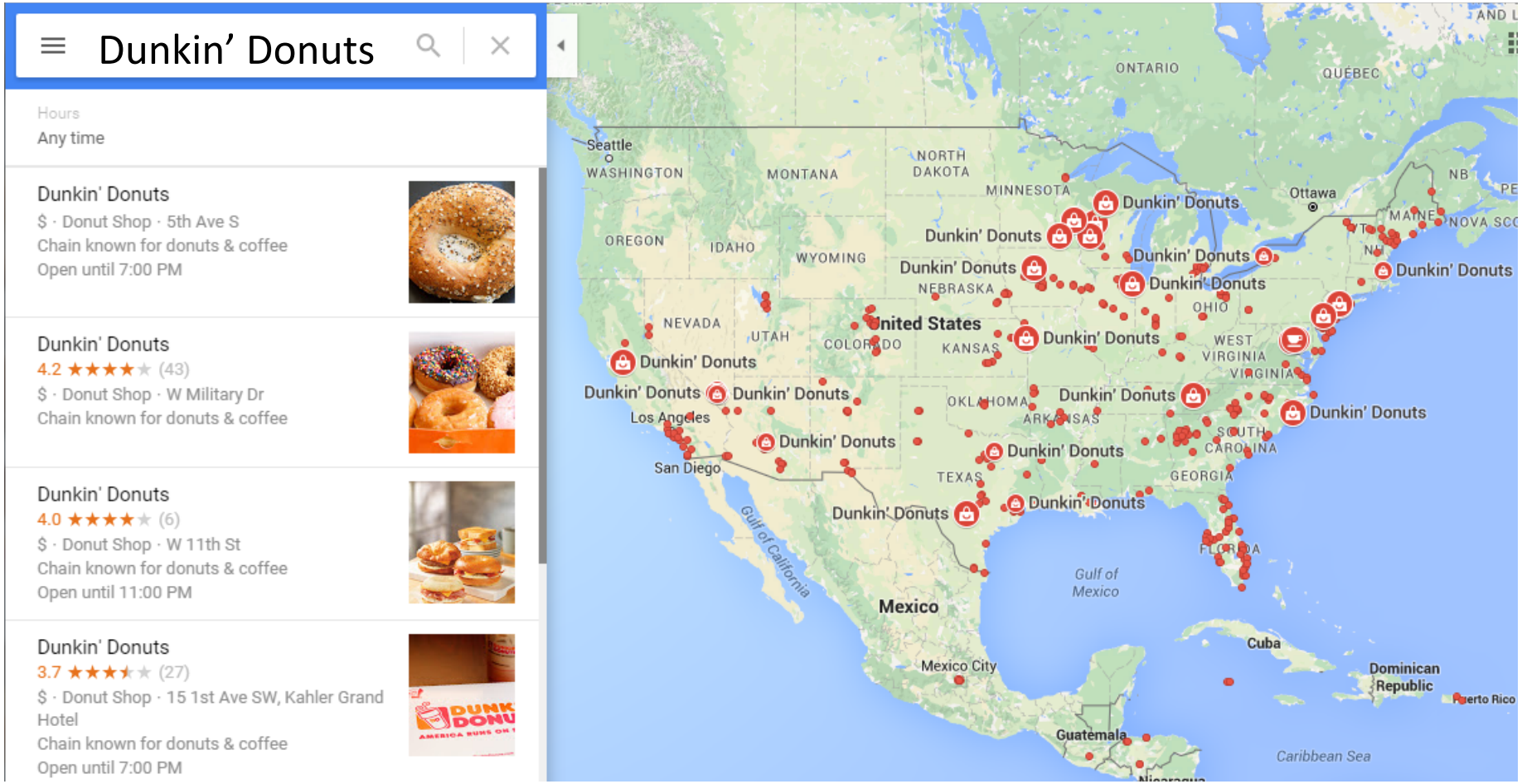


In collaboration with Judit Villen

DIA: Easy to collect, hard to unlock

- We want a query engine for DIA data
- We need and expect as good or better peptide detections than DDA
- We need to ensure no missing data across many runs
- Short processing times

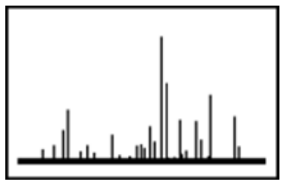
The Idea of a Query Engine for DIA data



Peptide-Centric Analysis as a Query Engine

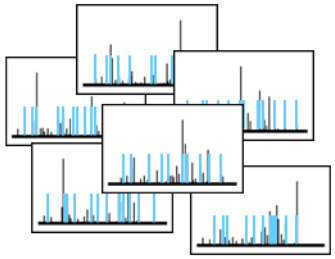
Spectrum-Centric Analysis

What peptide(s) best explain the data?



MS/MS spectrum

Protein sequence database



Peptide spectrum matches (PSMs)

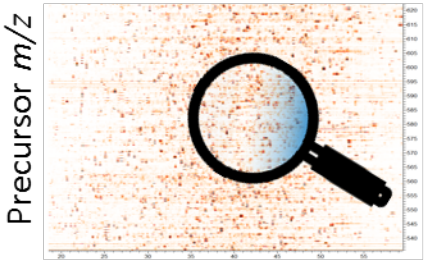
p-value
q-value

Peptide-Centric Analysis

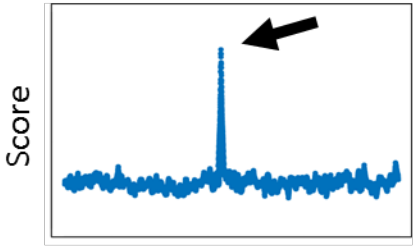
Is this peptide detected in the data?

NLPFSVENK

Peptide of interest



DIA data

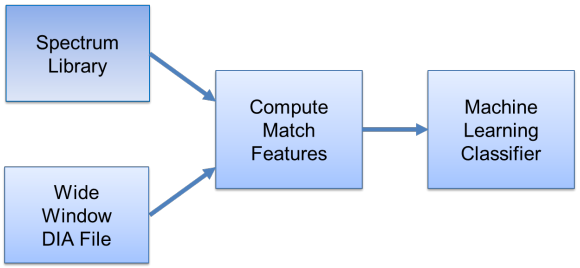


Evidence for detection

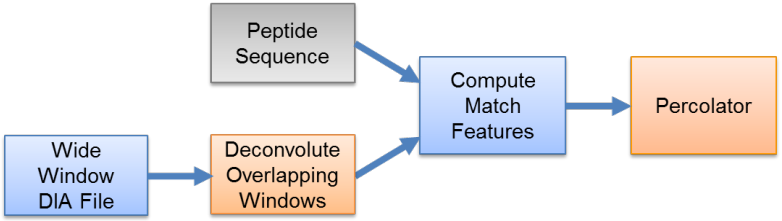
p-value
q-value

Two Innovative Tools for Peptide-Centric Analysis

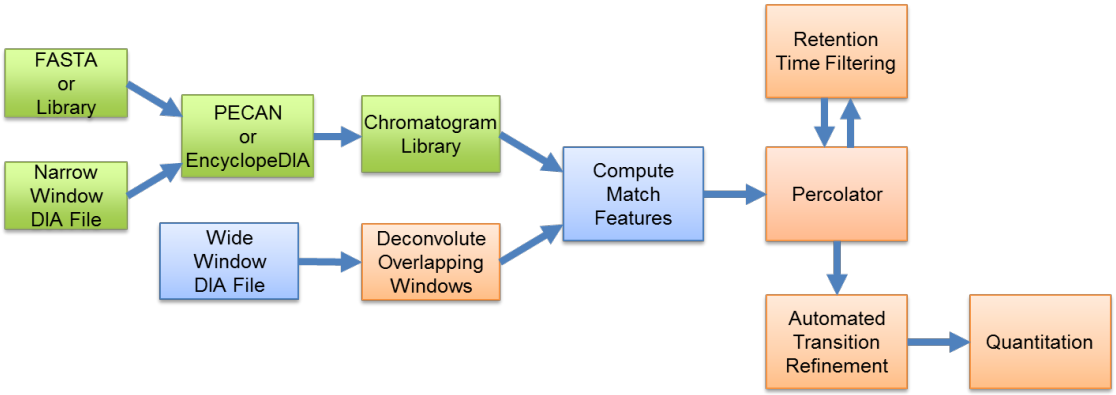
- Typical Workflow



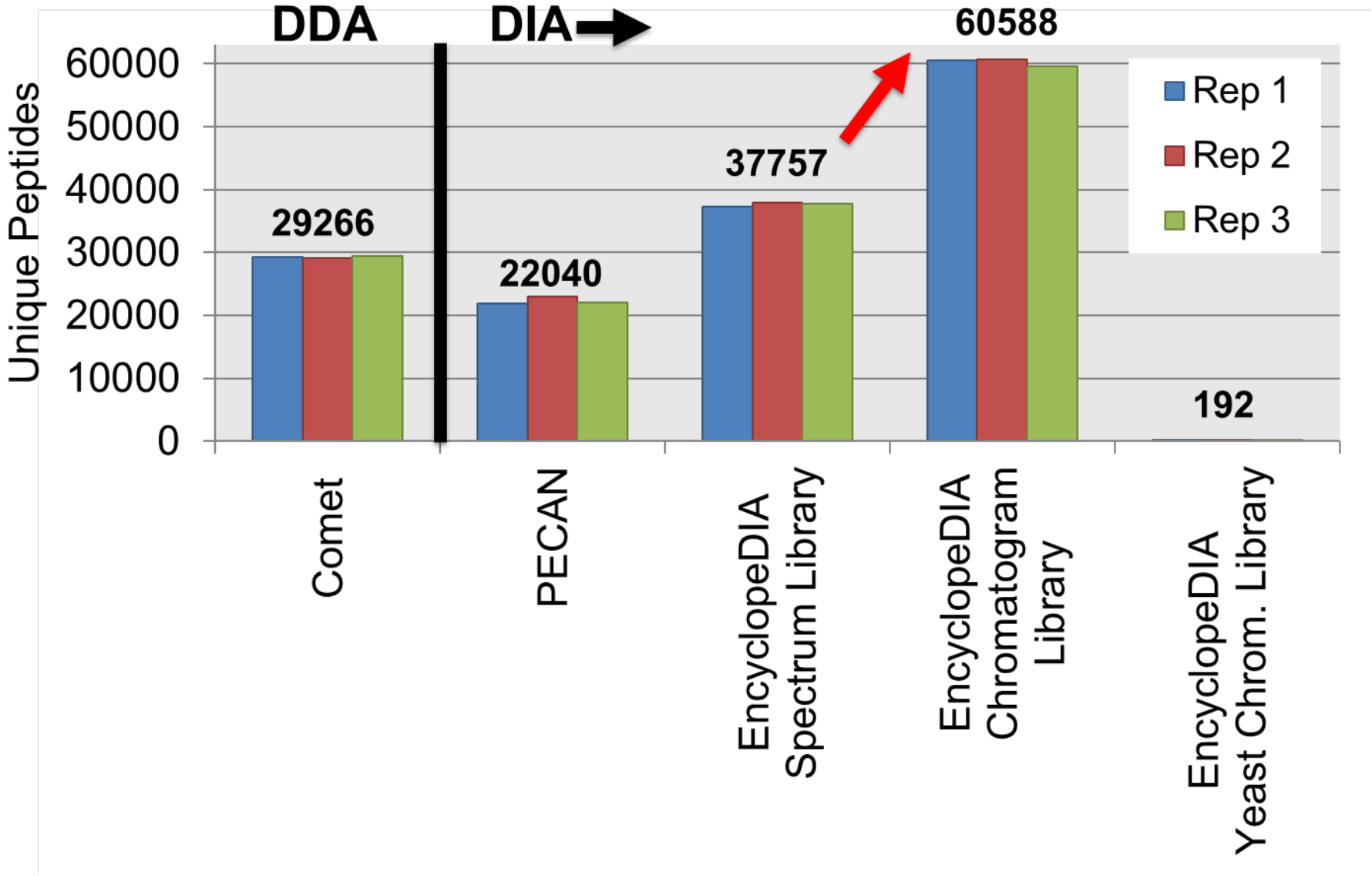
- Tool 1: Pecan
  - Uses only peptide sequences
  - No spectral or retention time information



- Tool 2: EncyclopeDIA
  - Uses peptide sequences
  - Spectrum or Chromatogram libraries
  - Positional isomers

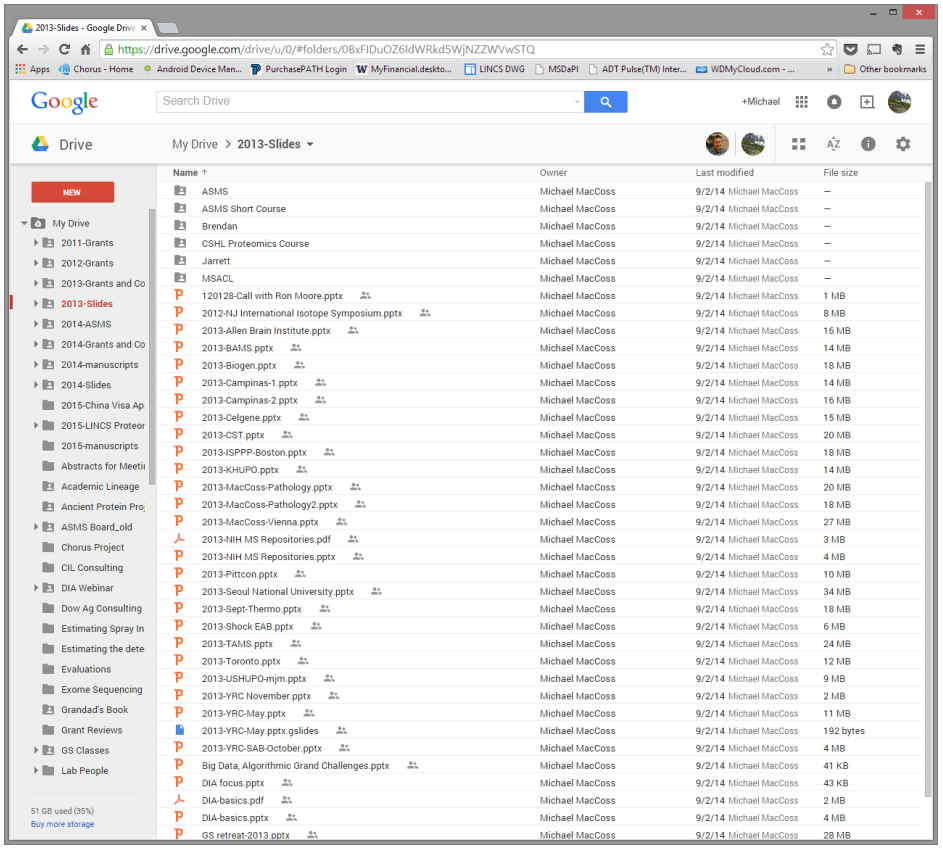


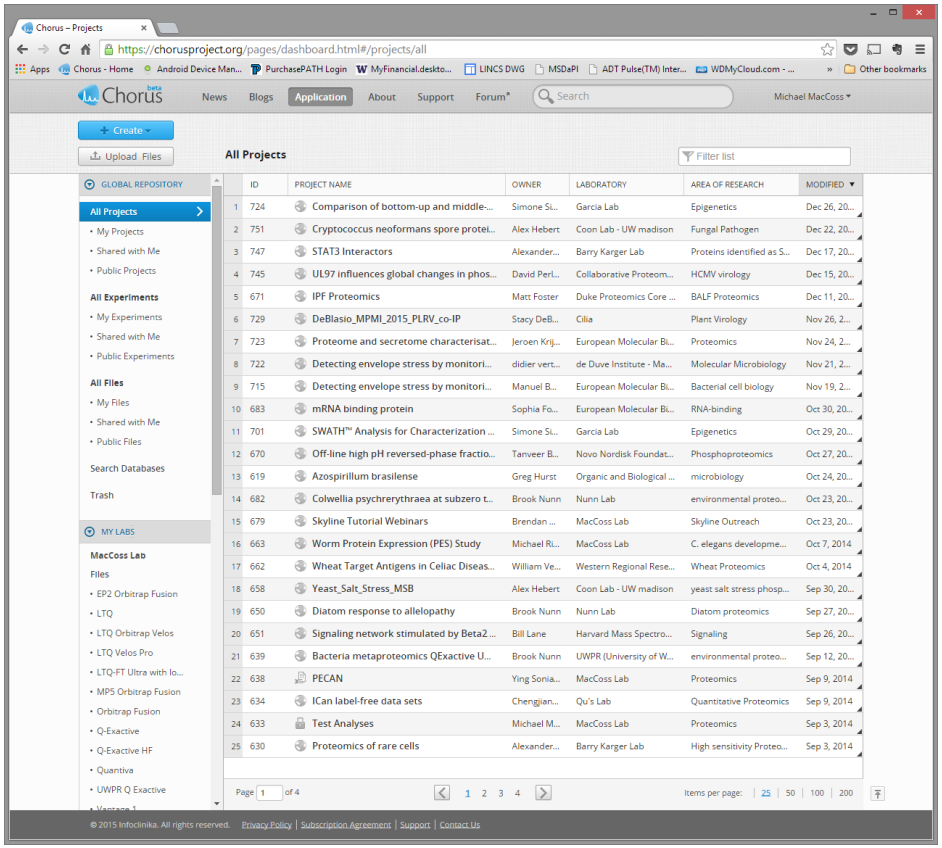
Pioneering methods meeting or exceeding goals



Chorus: Sharing and Dissemination of RAW Data

- The proteomics data needs to be “un-siloed”
  - A single place for all data
- Data is too large to move around.
- We need to bring algorithms to the data





Chorus allows extensive data organization

Chorus beta | News | Blogs | Application | About | Support | Forum | Search | Nathan Yates

Create | Upload Files

Public Projects | Filter list

GLOBAL REPOSITORY
- All Projects
- My Projects
- Shared with Me
- Public Projects
- All Experiments
- My Experiments
- Shared with Me
- Public Experiments
- All Files
- My Files
- Shared with Me
- Public Files
- Search Databases
- Trash

MY LABS | MY GROUPS | ADMINISTRATION TOOLS | REQUESTS

	ID	PROJECT NAME	OWNER	LABORATORY	AREA OF RESEARCH	MODIFIED
1	599	Bering Sea Sedimentary Proteins	Brook Nunn	UWPR (University of Washi...	environmental proteomics	Aug 18, 2014
2	438	Target Engagement Markers in Cerebrospinal Fluid - JPR ...	Nathan Yates	CHORUS Demo Lab	Proteomics	Aug 14, 2014
3	568	Arabidopsis WT and QR mutant Abscisic Acid Phosphopr...	Benjamin Mink...		Plant Proteomics	Jul 31, 2014
4	522	NeuCode SILAC	Anna Merrill	Coon Lab - UW madison	Proteomics	Jun 30, 2014
5	492	Fu				
6	457	St				
7	303	En				
8	366	Pr				
9	350	Al				
10	385	Al				
11	365	Hi				
12	351	Su				
13	352	Heterogeneous				
14	331	Growth Factor				
15	349	OSU_SIP				
16	345	hippocampi his				
17	325	Regulation of F				
18	134	WRN proteomic				
19	336	Disulfide bond				
20	96	Exhaustive imm				

Public Projects ›

NeuCode SILAC | Filter list

	ID	EXPERIMENT NAME	OWNER	LABORATORY	PROJECT	FILES
1	673	Multiplexing	Anna Merrill	Coon Lab - UW madison	NeuCode SILAC	14
2	672	NeuCode-TMT Comparison	Anna Merrill	Coon Lab - UW madison	NeuCode SILAC	24
3	671	Mixed Ratios	Anna Merrill	Coon Lab - UW madison	NeuCode SILAC	13

Public Projects › NeuCode SILAC ›

NeuCode-TMT Comparison | Filter

	ID	NAME	SIZE	INSTRUMENT	LABORATORY	UPLOAD DATE
1	39648	19Sep2013_4plexNeuCod...	667.91 MB	Charger	Coon Lab - UW madis...	Jun 30, 2014
2	39649	19Sep2013_4plexNeuCod...	654.67 MB	Charger	Coon Lab - UW madis...	Jun 30, 2014
3	39650	19Sep2013_4plexNeuCod...	637.12 MB	Charger	Coon Lab - UW madis...	Jun 30, 2014
4	39645	19Sep2013_4plexNeuCod...	742.07 MB	Charger	Coon Lab - UW madis...	Jun 30, 2014
5	39646	19Sep2013_4plexNeuCod...	700.27 MB	Charger	Coon Lab - UW madis...	Jun 30, 2014
6	39647	19Sep2013_4plexNeuCod...	675.21 MB	Charger	Coon Lab - UW madis...	Jun 30, 2014

Chorus to allow running pipelines in the cloud – nearly ready

- Database Search Engines
  - Comet (SEQUEST) and Percolator

537: With phosphorylation Processing Details

Create New Analysis

1 General Info

Analysis Name
My analysis

Workflow Type
Shotgun

Ratio
All Pairwise
Partitioning b
Factor
Time
Using
24

Description (optio

Create New Analysis

1 General Info

Isolation Width (min 0.5
2

Max Charge (min 1 - max
7

Rt Error Max
2

Mz Error Max
10

Protein Search Range
Overriding the default s
execution time. Howev

Min Rt

Min Mz

Cancel

Create New Analysis

1 General Info

Processi

Database
Mucata new

Precursor Mass Tolerance
20

Modification Parameters (Amino A
M 15.994 --...

Static Modification Parameters (A
C 57.021 --...

Mass Type Fragment
monoisotopic ma...

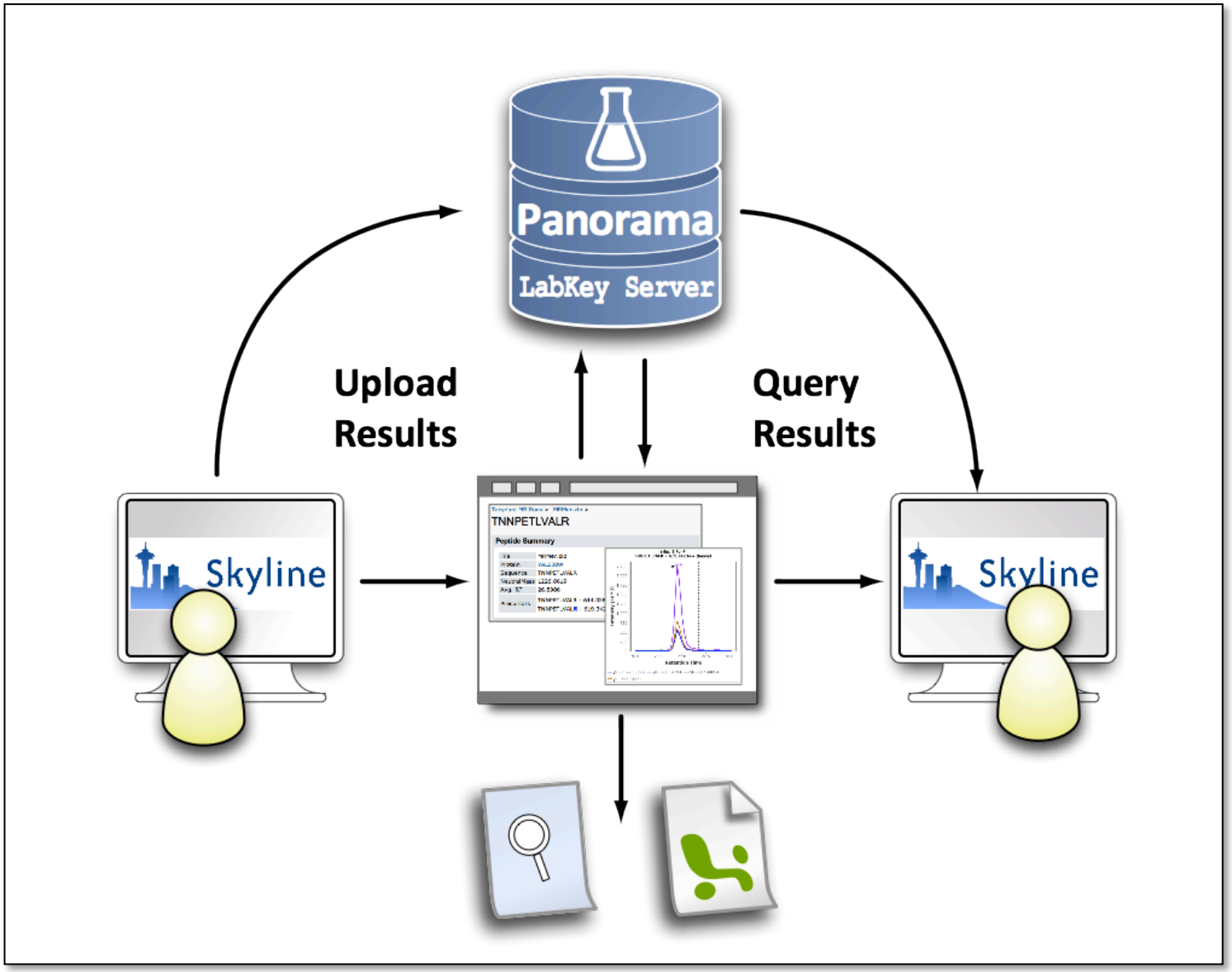
Create New Analysis

1 General Info | 2 Processing Params | 3 Comet Params | 4 Confirm

General	Name	My analysis
	Ratio	All pairwise
	Description	
	Workflow Type	Shotgun
Processing Params	Max Charge	7
	Isolation Width	2
	Search Bounds	Rt (from start - till end), Mz (from start - till end)
Comet Params	Database	Mucata new
	Mass Type Parent	monoisotopic masses
	Precursor Mass Tolerance	20
	Peptide Mass Units	ppm
	Dynamic Modifications	15.9949 M 0 3
	Static Modifications	57.021464 C
	Mass Type Fragment	monoisotopic masses

Cancel Back Run Analysis

Sharing processed data in Panorama



http://bit.ly/PCCSEData

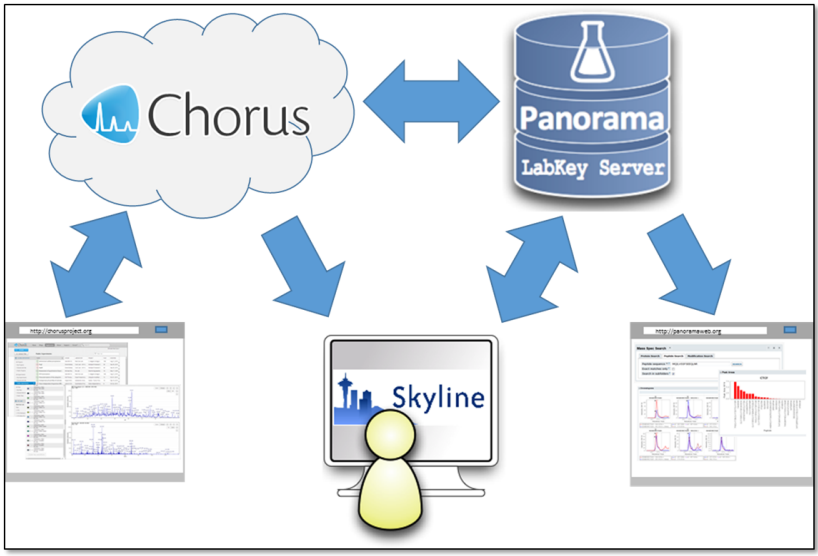
Next-Gen MS summary and outlook

- Comprehensive MS (**DIA**) is already here
  - Acquisition methods will continue to improve further

- We have **promising algorithms** to monitor unprecedented numbers of analytes in LINCS P100 DIA data

- Our Chorus cloud infrastructure will help us to **share** our data and promote **reuse**.

- Chorus will also enable use to efficiently **bring compute to the data** rather than vice versa.



PCCSE Overall Summary

- Great progress has been made in the first two years in establishing our center
- Our neurobiology models are progressing nicely and we are excited to extend these further
- Next-Gen MS holds great promise for increasing the impact of our work
- We are poised to use our data "as is" for comprehensive connectivity analysis, but also as a springboard for comparison with new data to be made via:
  - Our continued efforts
  - Our outreach and collaboration efforts
  - Data made by third parties