CSE 291: Biomolecular big data systems

Lecture 1: Introduction to the class, mass spectrometry and data repositories

Spring 2019

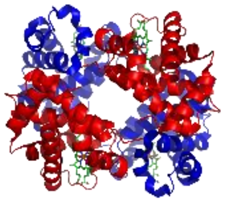
April 2, 2019

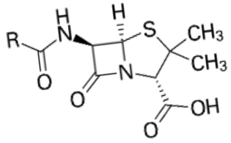
Center for
Computational
Mass
Spectrometry

UCSDCSE
Computer Science and Engineering

UC San Diego
SKAGGS SCHOOL OF PHARMACY
AND PHARMACEUTICAL SCIENCES

INTRODUCTION AND OVERVIEW





Course structure and approach

- Algorithms in the context of biomolecular big data
- Lectures cover algorithms but also hands-on data analysis on real datasets
- Lectures expected to be interactive with questions/discussion

Logistics: webpage, resources, staff, evaluation

Overview of the whole course

- Proteomics mass spectrometry
- High-throughput data analysis
- Computational systems used for the class
- NIH National Cancer Institute NCI60 cell lines

MassIVE

Mass Spectrometry
Interactive **V**irtual **E**nvironment

ProteoSAFe

COURSE RESOURCES AND STAFF

Course webpage:

http://proteomics.ucsd.edu/spring2019/cse291/

(laptop required for all lectures)

Resources

- Syllabus
- Lecture slides
- Homework assignments

Staff

Nuno Bandeira,
course director
bandeira@ucsd.edu
OH: Thu 3:30-4:50pm, CSE 4210

Benjamin Pullman, class TA
bpullman@eng.ucsd.edu
OH: Tue 3-5pm, CSE B275

TOPICS COVERED

MassIVE

Mass Spectrometry
Interactive Virtual Environment

ProteoSAFe

De novo sequencing; ab initio determination of peptide sequences directly from peptide spectra, greedy approaches, spectrum graphs, optimal dynamic programming approaches (including anti-symmetric path), generation of sequence tags.

Database search; scoring functions for matching spectra against peptide sequences, indexing strategies, grouping peptide identifications into protein identifications, false discovery rate corrections for multiple-hypothesis testing.

Spectral library searching; similarity scores for matching spectra, indexing and clustering strategies for matching large collections of spectra, false discovery rate corrections for multiple-hypothesis testing.

Post-translational modifications; defining the search space of all possible peptides with post-translational modifications (PTMs), Bayesian models of PTM-specific peptide fragmentation, sequence/spectrum alignment for blind discovery of unexpected PTMs, false localization rates for PTM site assignments.

Multi-spectrum identification; consensus interpretation of multiple spectra from peptides with overlapping sequences, spectrum/spectrum alignment, spectrum assembly (Overlap-Layout-Consensus and ABruijn graphs).

COURSE EVALUATION

Two quizzes, held in class (20% each, 40% of final grade)

- Tuesday, April 30th
- Thursday, June 6th

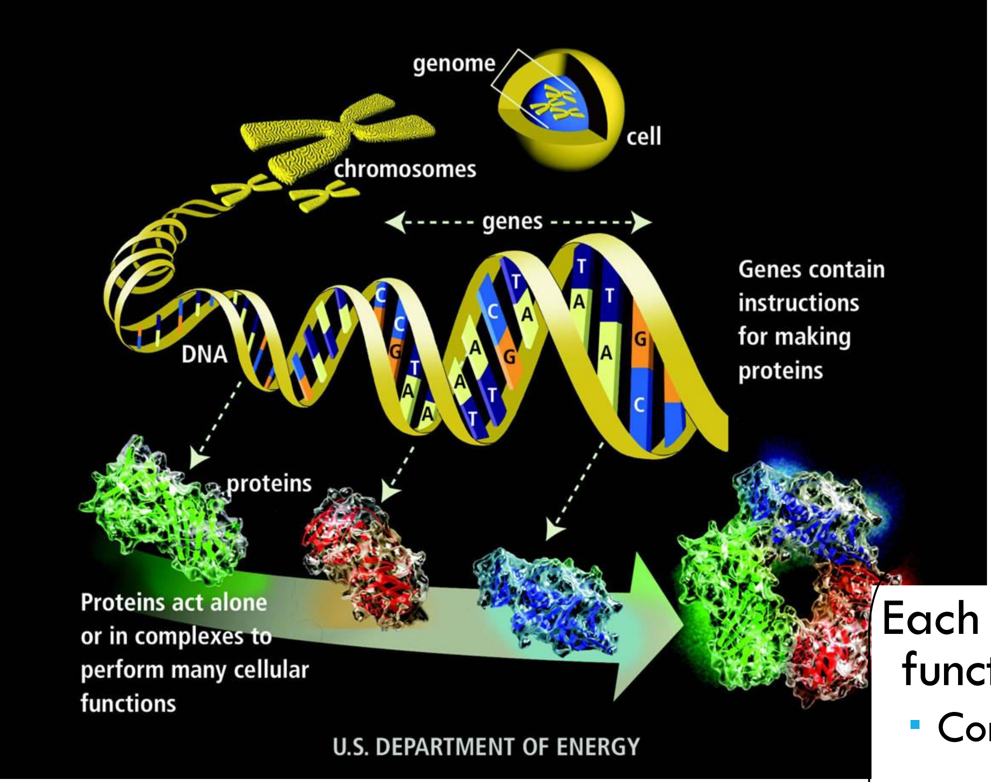
Homeworks (15% each, 30% of final grade)

- Tuesday, May 7th
- Thursday, June 6th

One group project (30% of final grade)

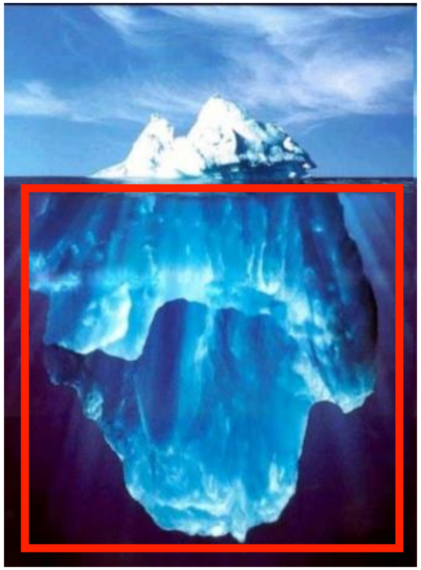
- Open-ended questions; will propose specific projects but exceptional project proposals may also be considered
- Project report and slides due June 10th
- Results presented during final exam time slot: June 11th, currently 8-11am (time might change)

PROTEINS RUN LIFE


genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. DEPARTMENT OF ENERGY

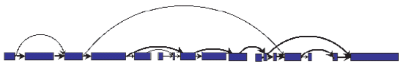
How much is in the genes?

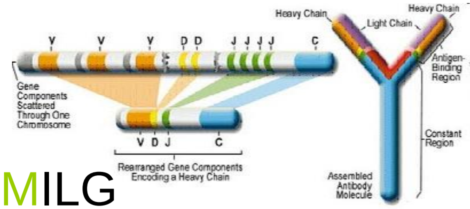
- Human: ~20,000-22,000
- Mouse: ~20,000-22,000
- Worm: ~19,000 (C.Elegans)
- Rice/Corn: ~32,000 – 45,000



Each gene may generate several proteins – the functional “workhorses” of the cell

- Combinatorial splicing




Heavy Chain
Light Chain
Heavy Chain
V
V
V
D
D
J
J
J
J
C
Antigen-Binding Region
Gene Components Scattered Through One Chromosome
V
D
J
C
Constant Region
Rearranged Gene Components Encoding a Heavy Chain
Assembled Antibody Molecule

- Sequence variation

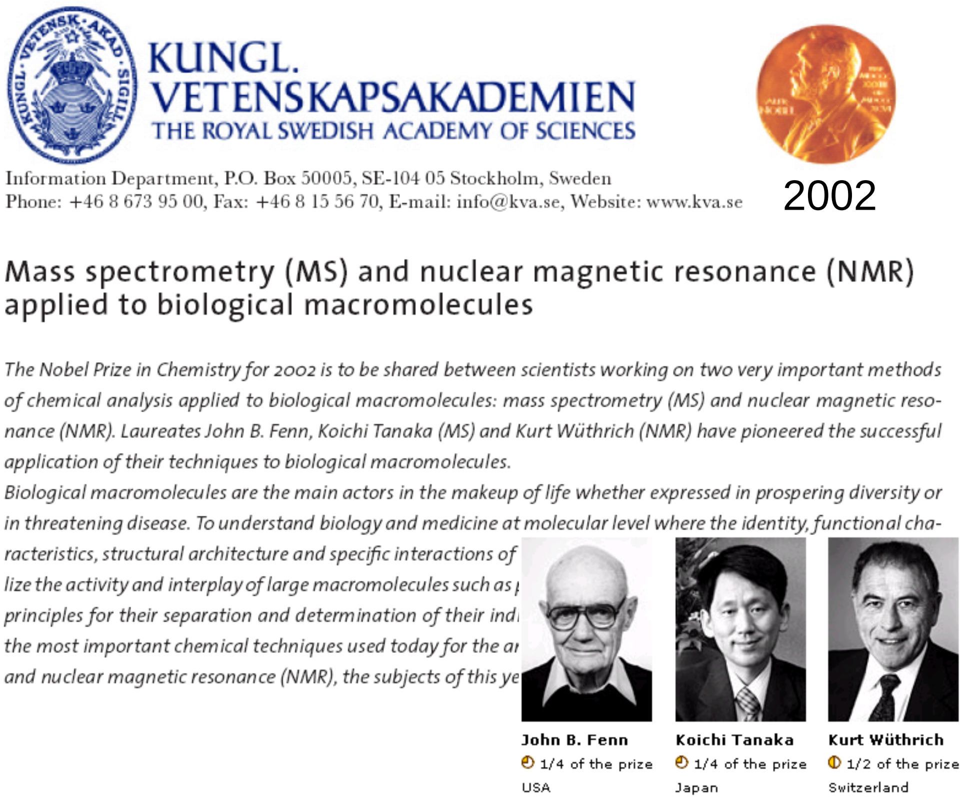
EMILG

EFILG

- Post-Translational Modifications
  - Hundreds of types and sites
- Quantification and turnover
- Protein Structure and interactions
- Endogenous and Immune peptides (e.g., insulin)
- Microbiome: 100-300x more genes

MASS SPECTROMETRY (MS)

Main technology for high-throughput analysis of proteins and small molecules



KUNGL. VETENSKAPSAKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

Information Department, P.O. Box 50005, SE-104 05 Stockholm, Sweden
Phone: +46 8 673 95 00, Fax: +46 8 15 56 70, E-mail: info@kva.se, Website: www.kva.se

2002

Mass spectrometry (MS) and nuclear magnetic resonance (NMR) applied to biological macromolecules

The Nobel Prize in Chemistry for 2002 is to be shared between scientists working on two very important methods of chemical analysis applied to biological macromolecules: mass spectrometry (MS) and nuclear magnetic resonance (NMR). Laureates John B. Fenn, Koichi Tanaka (MS) and Kurt Wüthrich (NMR) have pioneered the successful application of their techniques to biological macromolecules.

Biological macromolecules are the main actors in the makeup of life whether expressed in prospering diversity or in threatening disease. To understand biology and medicine at molecular level where the identity, functional characteristics, structural architecture and specific interactions of

lize the activity and interplay of large macromolecules such as

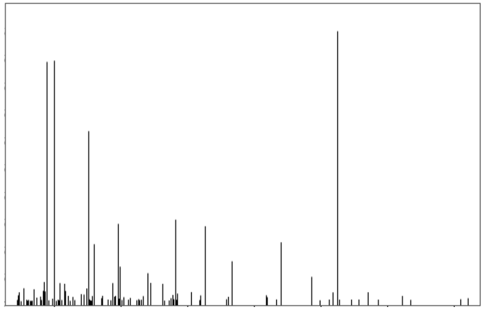
principles for their separation and determination of their ind

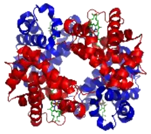
the most important chemical techniques used today for the a

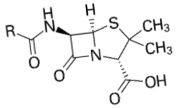
and nuclear magnetic resonance (NMR), the subjects of this ye

John B. Fenn	Koichi Tanaka	Kurt Wüthrich
1/4 of the prize	1/4 of the prize	1/2 of the prize
USA	Japan	Switzerland

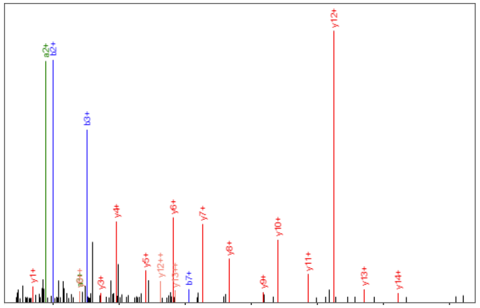
Fundamental computational question



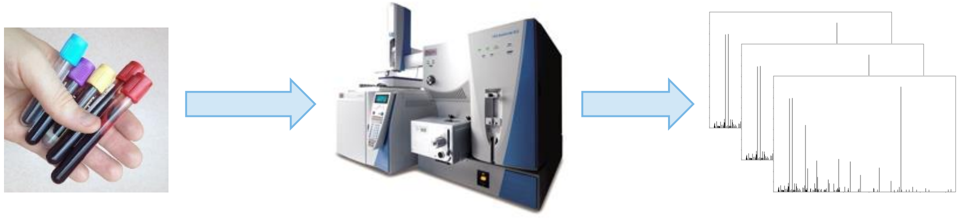




Protein?
Small molecule?
Modified?
Known or novel?



How to solve this problem billions of times for thousands of datasets?

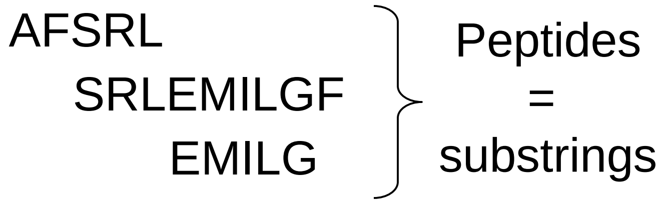


PROTEIN SEQUENCE

From a computational perspective, an amino acid sequence (protein or peptide) can be modeled as a string over a weighted alphabet:

Protein sequence:

…AFSRLEMILGF…



AFSRL

SRLEMILGF

EMILG

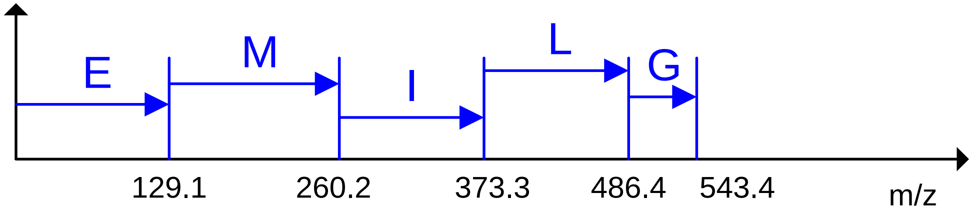
Peptides = substrings

Amino acid	Mass
A	71.0
F	147.1
S	87.0
R	156.1
L	113.1
E	129.1
M	131.1
I	113.1
G	57.0

WEIGHTED ALPHABET

Sequences of amino acids are almost equivalent to sequences of amino acid masses:

EMILG



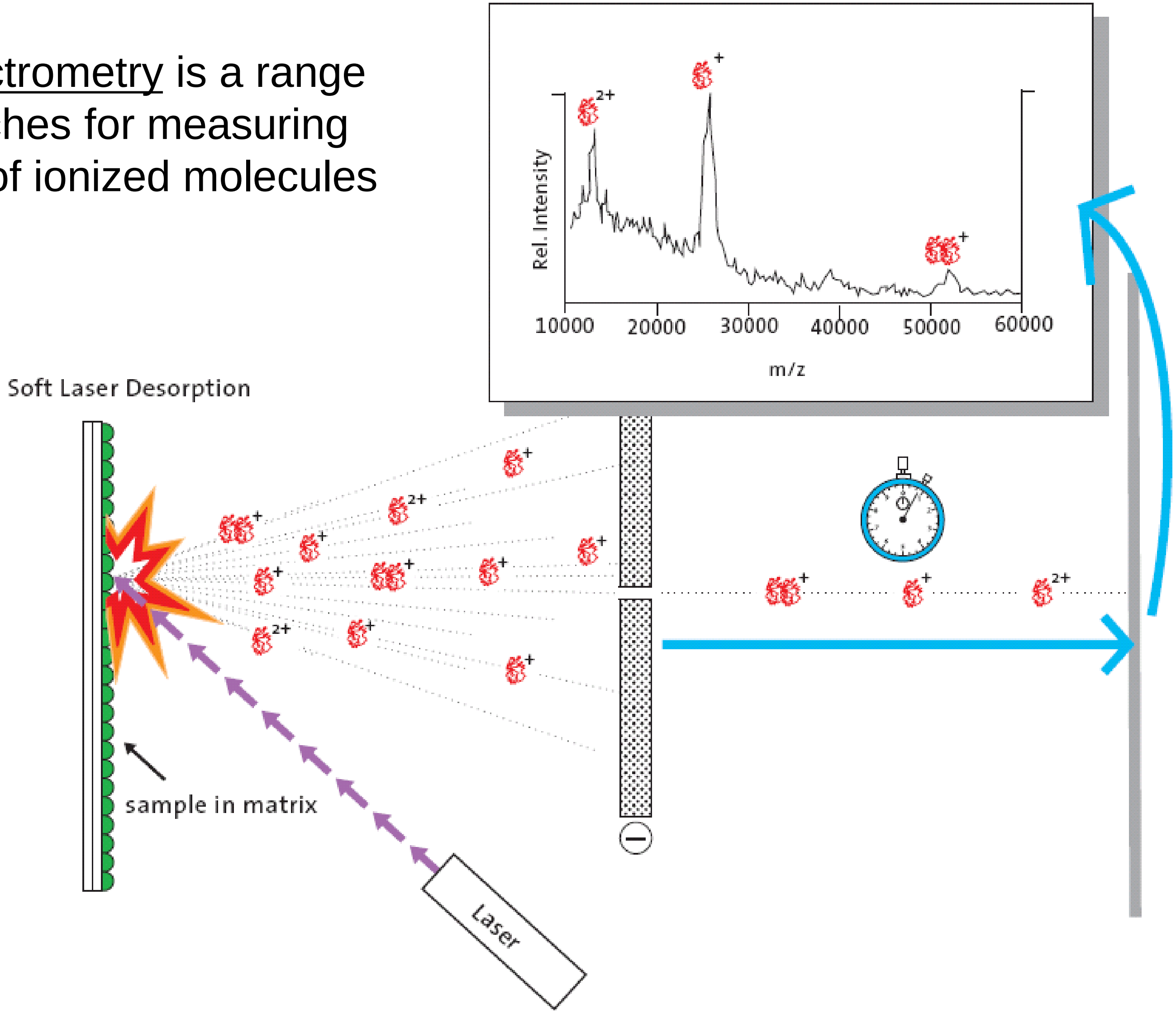
Exception:
m(I)=m(L)=113.1

Parent mass m(ρ) of a peptide $\rho=a_1,\ldots,a_n$ is given by $m(\rho)=\sum_{i=1..n}m(a_i)$

Mass spectrometry instruments allow us to measure the mass of molecules

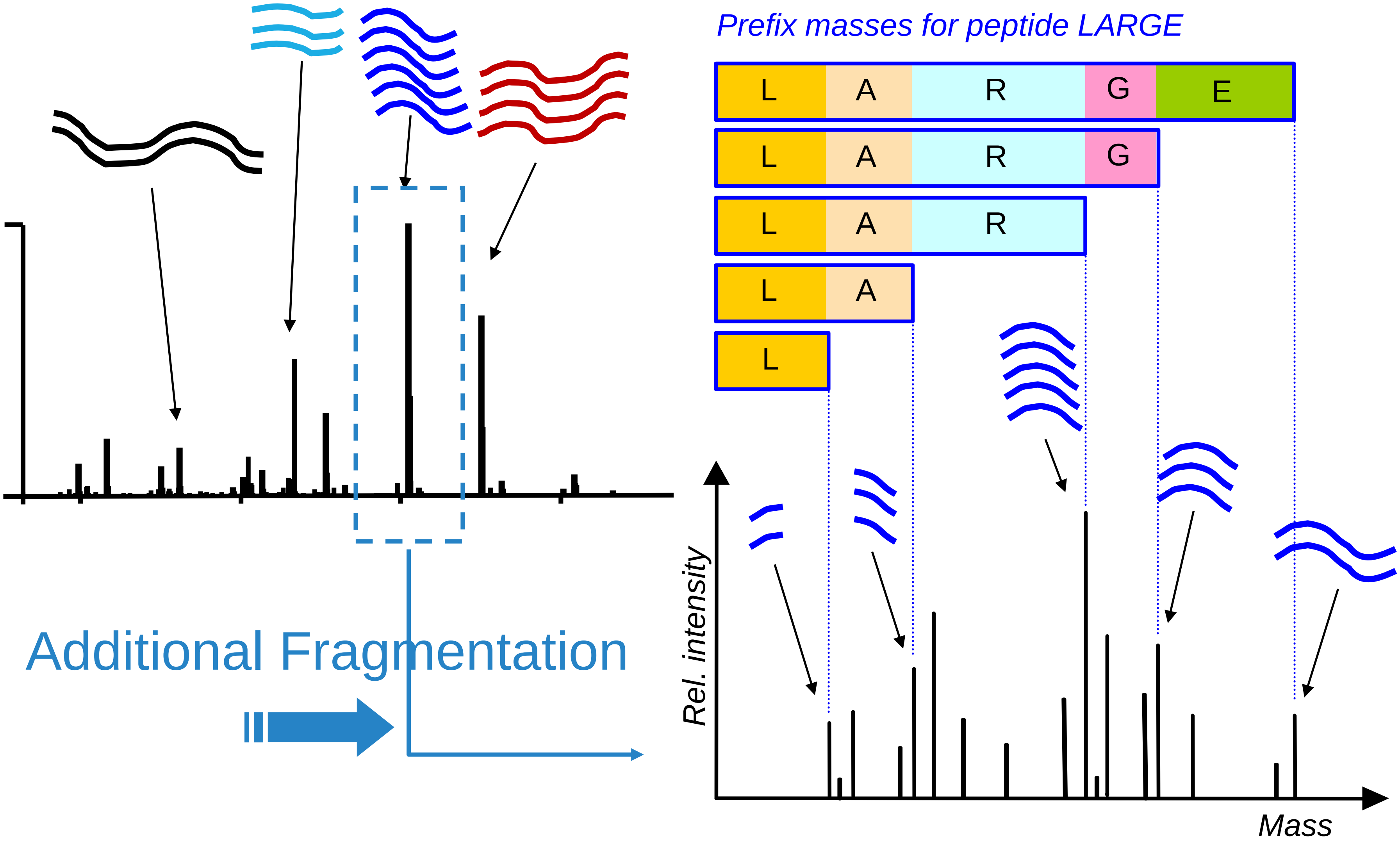
WHAT IS MASS SPECTROMETRY?

Mass spectrometry is a range of approaches for measuring the mass of ionized molecules



http://nobelprize.org/chemistry/laureates/2002/chemadv02.pdf

TANDEM MASS SPECTROMETRY (MS/MS)



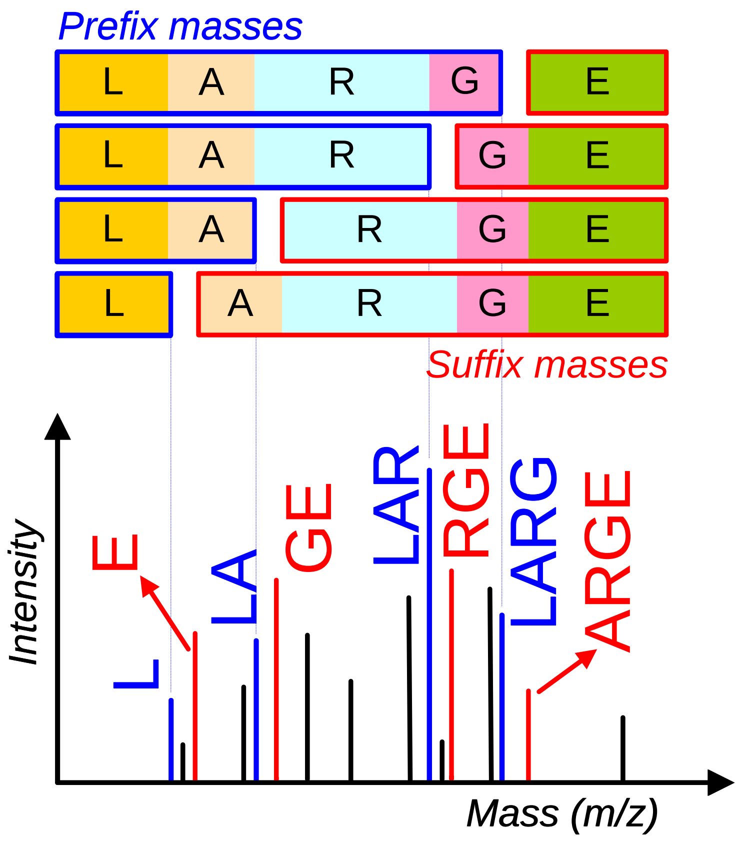
Additional Fragmentation

Prefix masses for peptide LARGE

Adapted from slides by Vineet Bafna, UCSD

Tandem Mass Spectrometry (MS/MS)

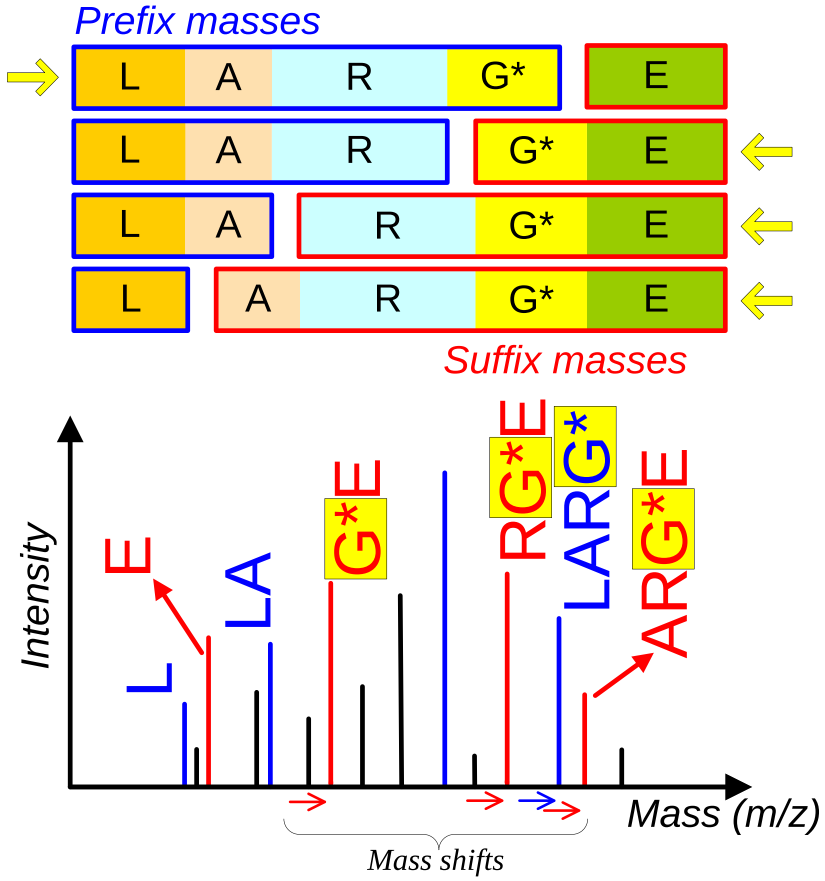
Peptide LARGE



MS/MS spectrum

Modified peptide
LARG*E

Modification: any event that changes the mass at a specific site.

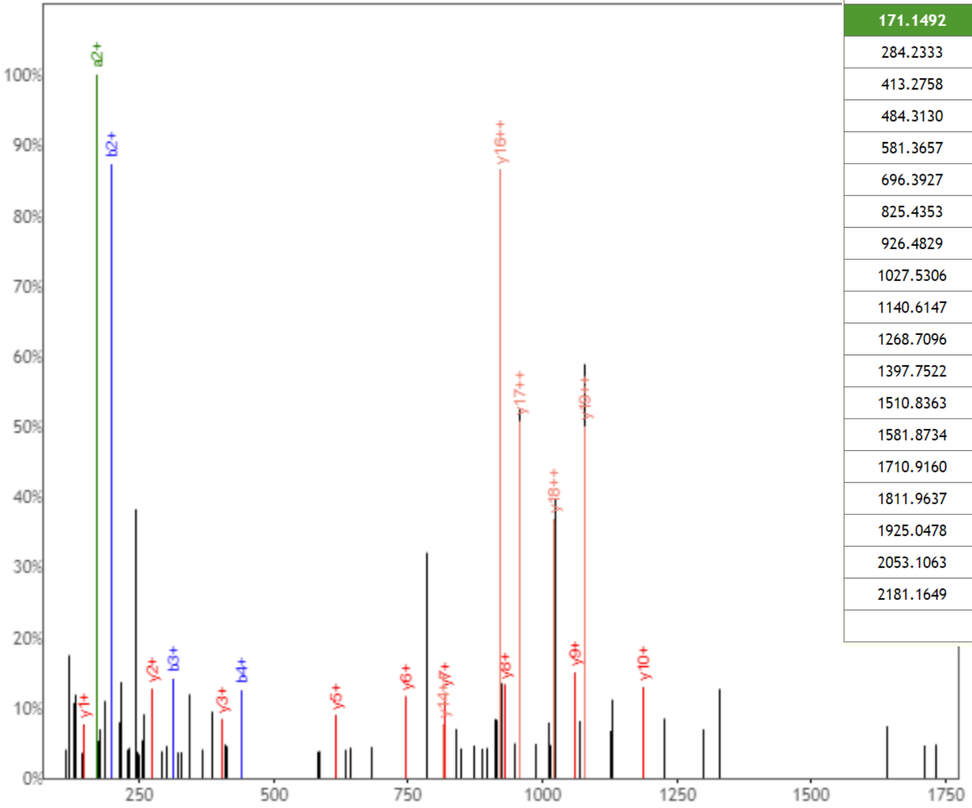


EXAMPLE OF A REAL MS/MS SPECTRUM

Peptide VVLEAPDETTLKELAETLQQK, MH+ 2355.2653, charge 3 [interactive view]

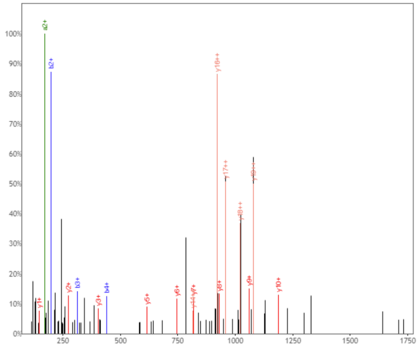
a+	b+	#	Seq	#	y+	y2+
72.0808	100.0757	1	V	21		
171.1492	**199.1441**	2	V	20	2256.1969	1128.6021
284.2333	**312.2282**	3	L	19	2157.1285	**1079.0679**
413.2758	**441.2708**	4	E	18	2044.0445	**1022.5259**
484.3130	512.3079	5	A	17	1915.0019	**958.0046**
581.3657	609.3606	6	P	16	1843.9647	**922.4860**
696.3927	724.3876	7	D	15	1746.9120	873.9596
825.4353	853.4302	8	E	14	1631.8850	**816.4462**
926.4829	954.4779	9	T	13	1502.8424	751.9249
1027.5306	1055.5255	10	T	12	1401.7948	701.4010
1140.6147	1168.6096	11	L	11	1300.7471	650.8772
1268.7096	1296.7046	12	K	10	**1187.6630**	594.3352
1397.7522	1425.7472	13	E	9	**1059.5681**	530.2877
1510.8363	1538.8312	14	L	8	**930.5255**	465.7664
1581.8734	1609.8683	15	A	7	**817.4414**	409.2243
1710.9160	1738.9109	16	E	6	**746.4043**	373.7058
1811.9637	1839.9586	17	T	5	**617.3617**	309.1845
1925.0478	1953.0427	18	L	4	516.3140	258.6606
2053.1063	2081.1012	19	Q	3	**403.2300**	202.1186
2181.1649	2209.1598	20	Q	2	**275.1714**	138.0893
		21	K	1	**147.1128**	74.0600

Fragment ions mass table



Annotated MS/MS spectrum

PEPTIDE-SPECTRUM MATCH (PSM)



Annotation of spectrum peaks with peptide fragment ions

- Sequence prefix fragments generate (mostly) b ions, a ions are also often prominent
- Sequence suffix fragments generate (mostly) y ions
- Ions have to be charged to be detectable by mass spectrometry
  - Can sometimes be multiply-charged
  - ^{13}C isotopes help determine fragment charges

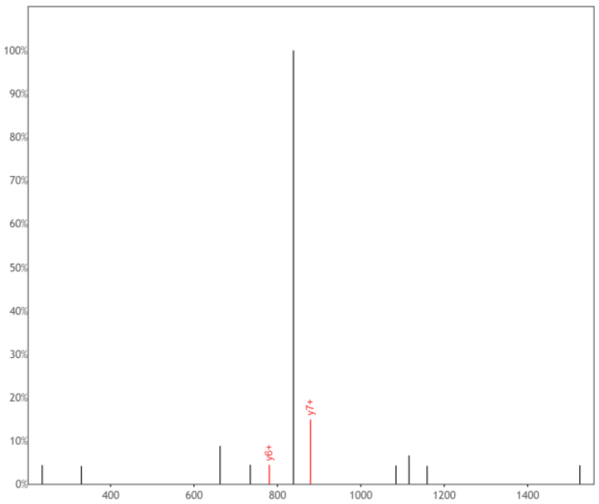
Good peptide-spectrum matches usually have the following features

- High percentage of total intensity is explained by peptide ions at low mass tolerance
- Consecutive series of y fragments (sometimes also of b ions)
- Known sequence-specific fragmentation patterns; e.g., high-intensity peaks before/after amino acid P (Proline)

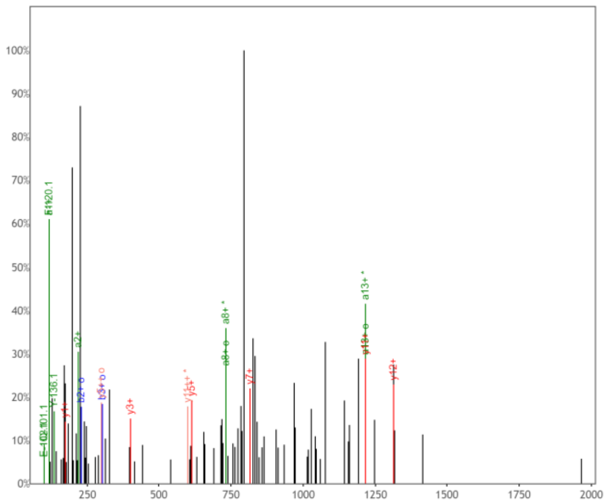
POOR PEPTIDE SPECTRUM MATCHES

- HDSKWFKEPYFVHAVEWGSHVYFFFR: insufficient matched peaks; identification mostly based on absence of better alternative explanations
- FTASAGIQVVGDDLTVTNPKR: many unexplained peaks, low explained intensity low signal-to-noise ratio

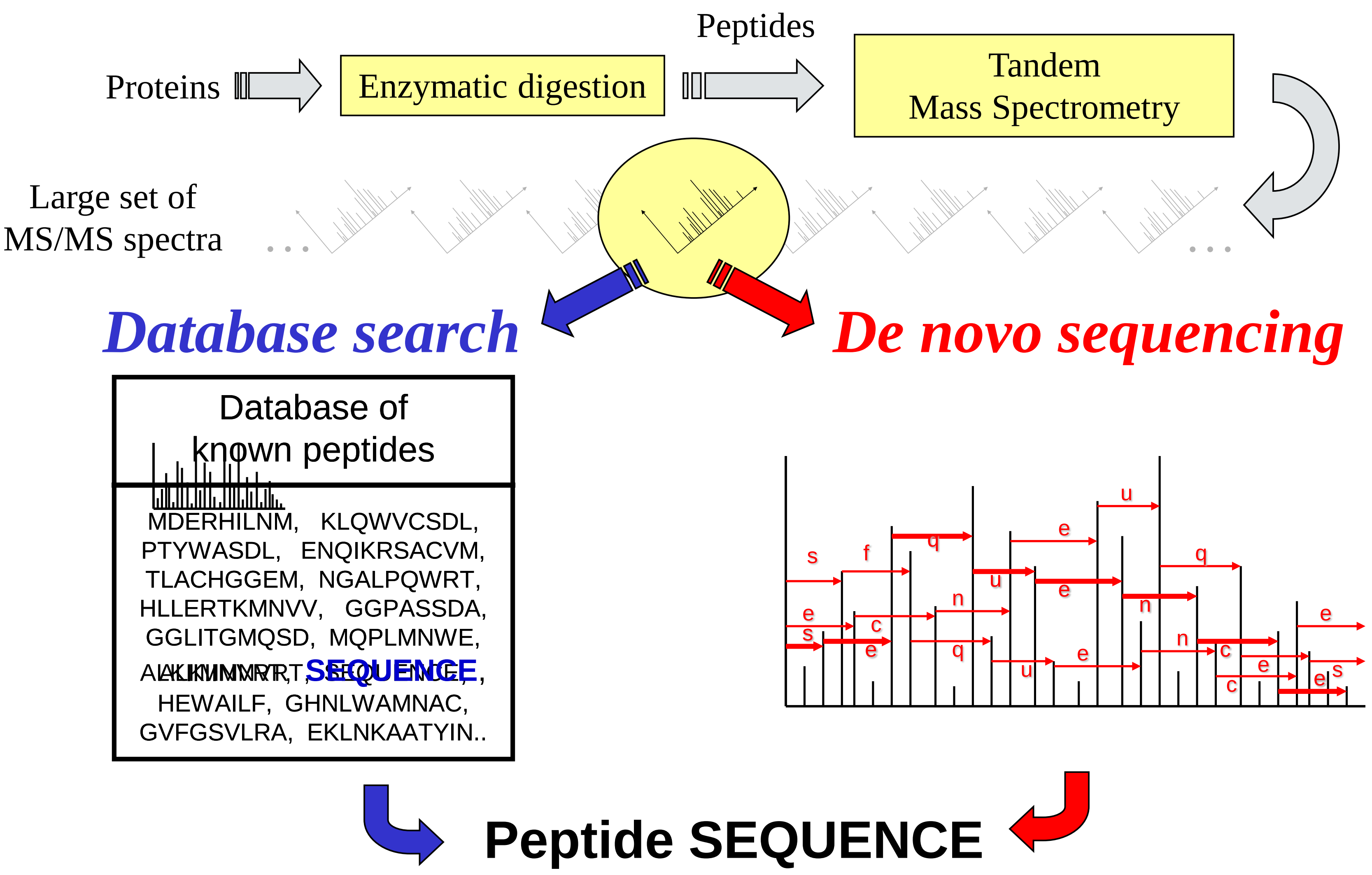
HDSKWFKEPYFVHAVEWGSHVYFFFR



FTASAGIQVVGDDLTVTNPKR



TANDEM MASS SPECTROMETRY (MS/MS) IDENTIFICATION



Proteins

Enzymatic digestion

Peptides

Tandem
Mass Spectrometry

Large set of
MS/MS spectra

. . .

. . .

Database search

De novo sequencing

Database of
known peptides

MDERHILNM, KLQWVCSDL,
PTYWASDL, ENQIKRSACVM,
TLACHGGEM, NGALPQWRT,
HLLERTKMNVV, GGPASSDA,
GGLITGMQSD, MQPLMNWE,
ALKIMMRRT, **SEQUENCE**,
HEWAILF, GHNLWAMNAC,
GVFGSVLRA, EKLNKAATYIN..

Peptide SEQUENCE

DATABASE SEARCH VS DE NOVO

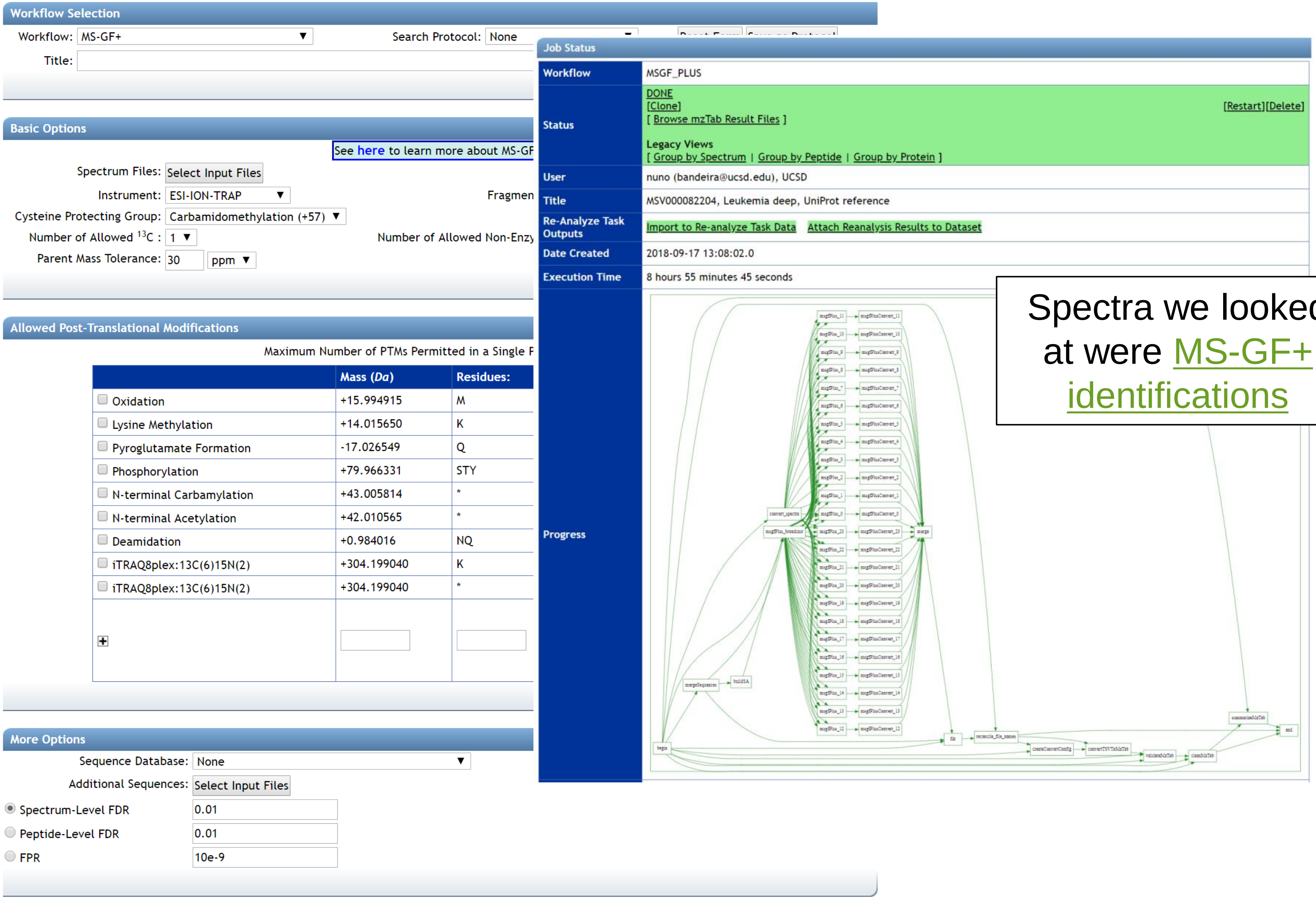
Database search

- Pro: restricts the search space to sequences with higher likelihood of being correct (e.g., derived from the genome)
- Pro: smaller search space increases the contrast between possible alternative explanations
- Pro: applicable to a much larger fraction of mass spectrometry data
- Con: not applicable to novel proteins such as antibodies and cancer proteins
- Con: hard to predict unexpected mutations/modifications and highly modified peptides

De novo sequencing

- Pro: does not require a database of protein sequences
- Pro: can determine sequences that have never been seen before
- Con: requires very high quality spectra with extensive fragmentation and low noise levels
- Con: sequencing error rates can be as high as one error per 4 predicted amino acids
- Con: tends to generate only short sequences

MS-GF+ DATABASE SEARCH (WORKFLOW HERE)



Workflow Selection

Workflow: MS-GF+ ▼

Search Protocol: None

Title:

Basic Options

See here to learn more about MS-GF

Spectrum Files: Select Input Files

Instrument: ESI-ION-TRAP ▼

Fragmen

Cysteine Protecting Group: Carbamidomethylation (+57) ▼

Number of Allowed ^{13}C : 1 ▼

Number of Allowed Non-Enzy

Parent Mass Tolerance: 30 ppm ▼

Allowed Post-Translational Modifications

Maximum Number of PTMs Permitted in a Single P

	Mass (*Da*)	Residues:
□ Oxidation	+15.994915	M
□ Lysine Methylation	+14.015650	K
□ Pyroglutamate Formation	-17.026549	Q
□ Phosphorylation	+79.966331	STY
□ N-terminal Carbamylation	+43.005814	*
□ N-terminal Acetylation	+42.010565	*
□ Deamidation	+0.984016	NQ
□ iTRAQ8plex:13C(6)15N(2)	+304.199040	K
□ iTRAQ8plex:13C(6)15N(2)	+304.199040	*
⊞		

More Options

Sequence Database: None ▼

Additional Sequences: Select Input Files

● Spectrum-Level FDR 0.01

○ Peptide-Level FDR 0.01

○ FPR 10e-9

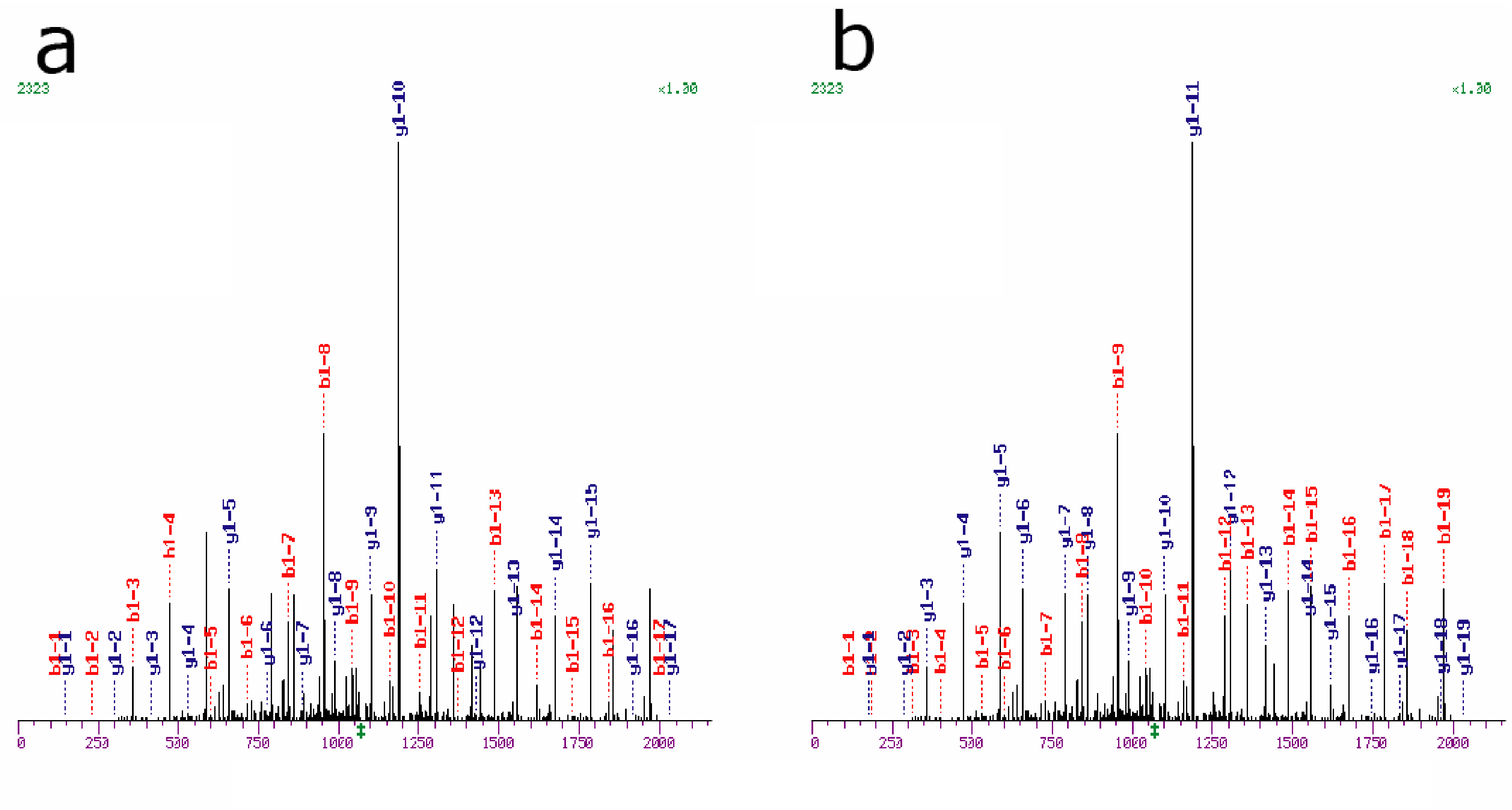
Job Status

Workflow	MSGF_PLUS
Status	DONE [Clone] [Browse mzTab Result Files] Legacy Views [Group by Spectrum | Group by Peptide | Group by Protein] [Restart][Delete]
User	nuno (bandeira@ucsd.edu), UCSD
Title	MSV000082204, Leukemia deep, UniProt reference
Re-Analyze Task Outputs	Import to Re-analyze Task Data Attach Reanalysis Results to Dataset
Date Created	2018-09-17 13:08:02.0
Execution Time	8 hours 55 minutes 45 seconds
Progress	

Spectra we looked at were MS-GF+ identifications

DETERMINING RELIABILITY OF IDENTIFICATIONS

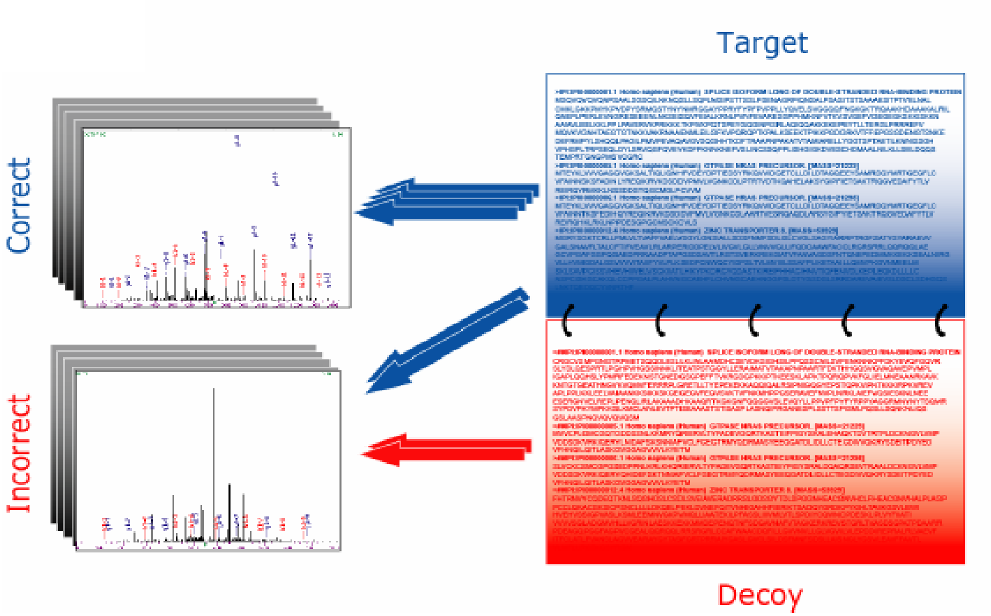
EVERY spectrum has some best match to the database – how can we tell whether it's a significant match?



Elias & Gygi '07

TARGET/DECOY APPROACH (TDA)

Decoy databases are the most common approach to determine the reliability of identifications – estimate the False Discovery Rate (FDR)



Null hypothesis of the target-decoy strategy:

- Each spectrum is generated by a random (peptide-like) amino acid sequence
- Number of false matches to target equals number of matches to decoy
- FDR defined as #decoy_matches / #target_matches

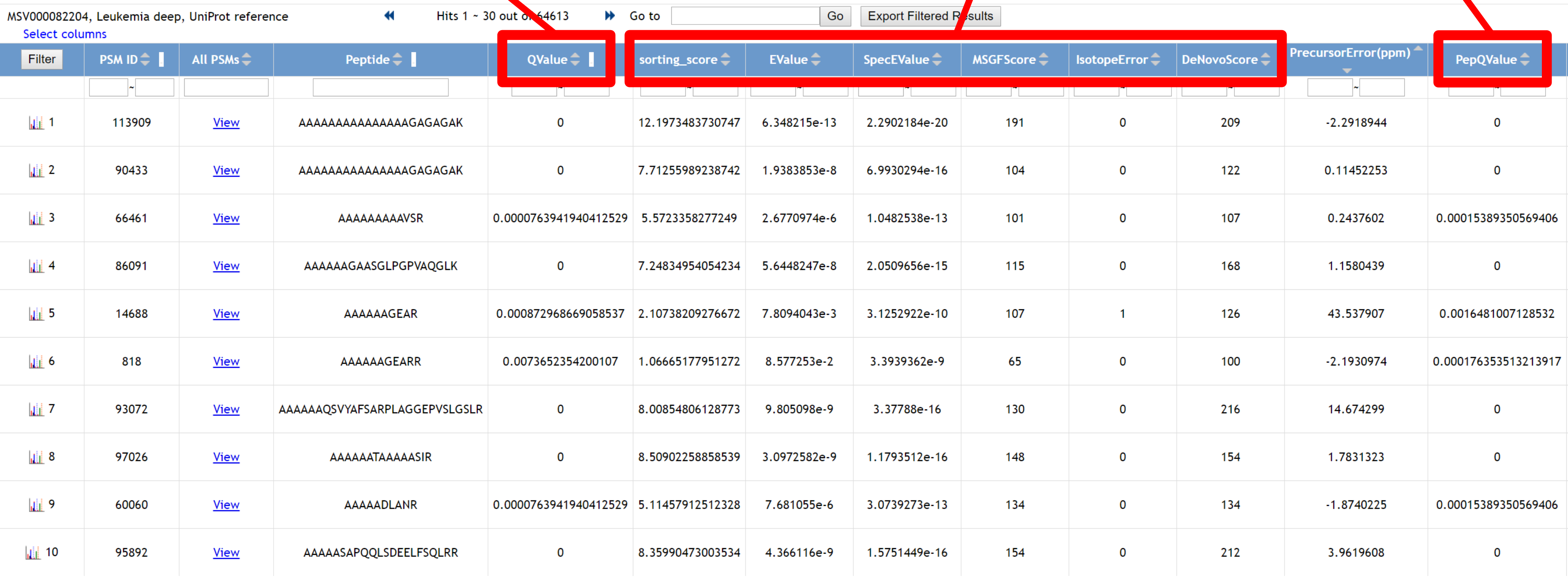
Elias and Gygi '07

FDR REPORTED IN RESULTS VIEWS AS Q-VALUE

Q-value

Match scores

Peptide Q-value



MSV000082204, Leukemia deep, UniProt reference

Select columns

Hits 1 ~ 30 out of 64613 Go to Go Export Filtered Results

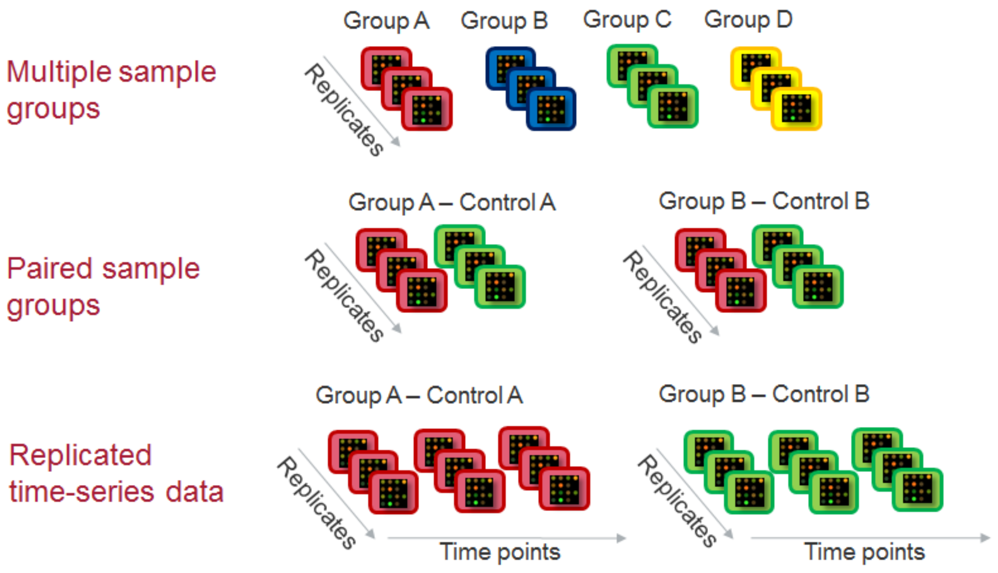
Filter	PSM ID	All PSMs	Peptide	QValue	sorting_score	EValue	SpecEValue	MSGFScore	IsotopeError	DeNovoScore	PrecursorError(ppm)	PepQValue
1	113909	View	AAAAAAAAAAAAAAAAGAGAGAK	0	12.1973483730747	6.348215e-13	2.2902184e-20	191	0	209	-2.2918944	0
2	90433	View	AAAAAAAAAAAAAAAAGAGAGAK	0	7.71255989238742	1.9383853e-8	6.9930294e-16	104	0	122	0.11452253	0
3	66461	View	AAAAAAAAAAVSR	0.0000763941940412529	5.5723358277249	2.6770974e-6	1.0482538e-13	101	0	107	0.2437602	0.00015389350569406
4	86091	View	AAAAAAGAASGLPGPVAQGLK	0	7.24834954054234	5.6448247e-8	2.0509656e-15	115	0	168	1.1580439	0
5	14688	View	AAAAAAGEAR	0.000872968669058537	2.10738209276672	7.8094043e-3	3.1252922e-10	107	1	126	43.537907	0.0016481007128532
6	818	View	AAAAAAGEARR	0.0073652354200107	1.06665177951272	8.577253e-2	3.3939362e-9	65	0	100	-2.1930974	0.000176353513213917
7	93072	View	AAAAAAQSVYAFSARPLAGGEPVSLGSLR	0	8.00854806128773	9.805098e-9	3.37788e-16	130	0	216	14.674299	0
8	97026	View	AAAAAATAAAAASIR	0	8.50902258858539	3.0972582e-9	1.1793512e-16	148	0	154	1.7831323	0
9	60060	View	AAAAADLANR	0.0000763941940412529	5.11457912512328	7.681055e-6	3.0739273e-13	134	0	134	-1.8740225	0.00015389350569406
10	95892	View	AAAAASAPQQLSDEELFSQLRR	0	8.35990473003534	4.366116e-9	1.5751449e-16	154	0	212	3.9619608	0

(link to search results)

EXPERIMENTAL DESIGN METADATA

Datasets have metadata defining sample types and conditions

- Conditions or groups: typically healthy-vs-disease
- Biological replicates: different samples, typically different individuals; main aim is to average biological variation unrelated to conditions of interest
- Technical replicates: repeated experimental runs of the same sample; main aim is to average out technical variability (e.g., `noise')



(figure reused from http://www.ra.cs.uni-tuebingen.de/software/RPPApipe/doc/documentation.htm)

METADATA ENABLES QUANTITATIVE ANALYSIS

Simplest design considers two conditions with multiple biological replicates per group:

1. Measure abundance of analytes per group

- Analytes can be proteins, small molecules, peptides, drug byproducts, etc
- Abundance is measured in proportion to the number of analyte molecules in each group – can be counts of spectra identified to analyte or can be total intensity of ions assigned to analyte

2. Determine changes in abundance across the conditions of interest (differential expression)

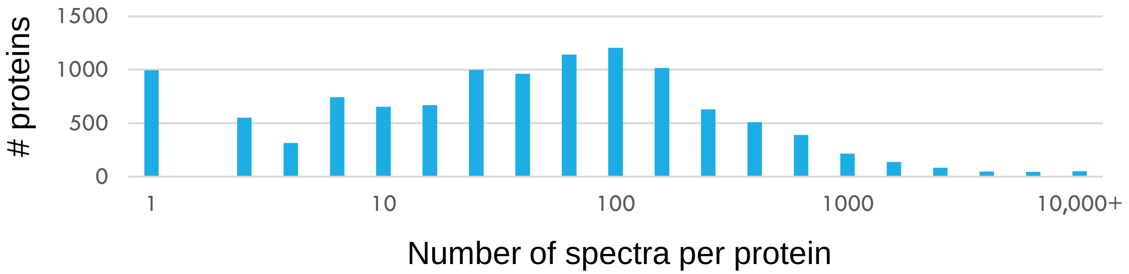
3. Assess the statistical significance of observed changes

- Naively summing abundance per group ignores intra-group variation
- Need to consider the consistency of abundances within each group
- Typical statistical tests for comparison of means are t-test for Gaussian distributions and Mann-Whitney U-test for non-Gaussian distributions
  - Null hypothesis is that the two distributions have the same mean

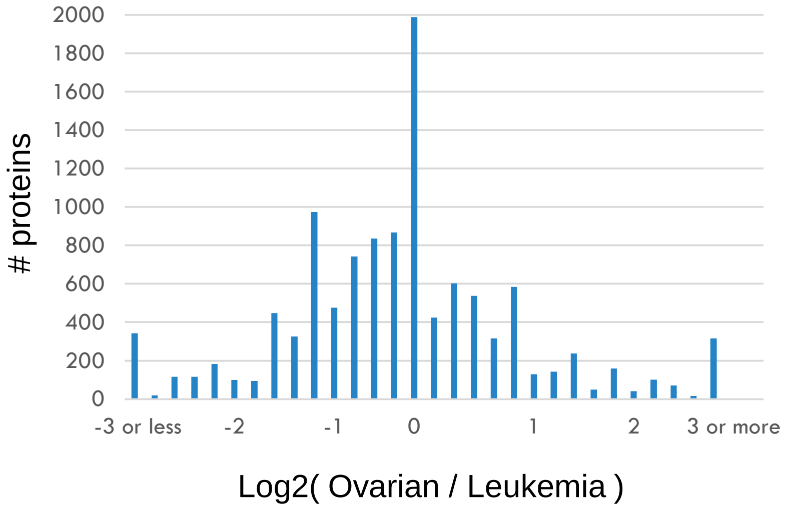
DIFFERENTIAL EXPRESSION BETWEEN TISSUES

Leukemia cell line compared to ovarian cancer cell line

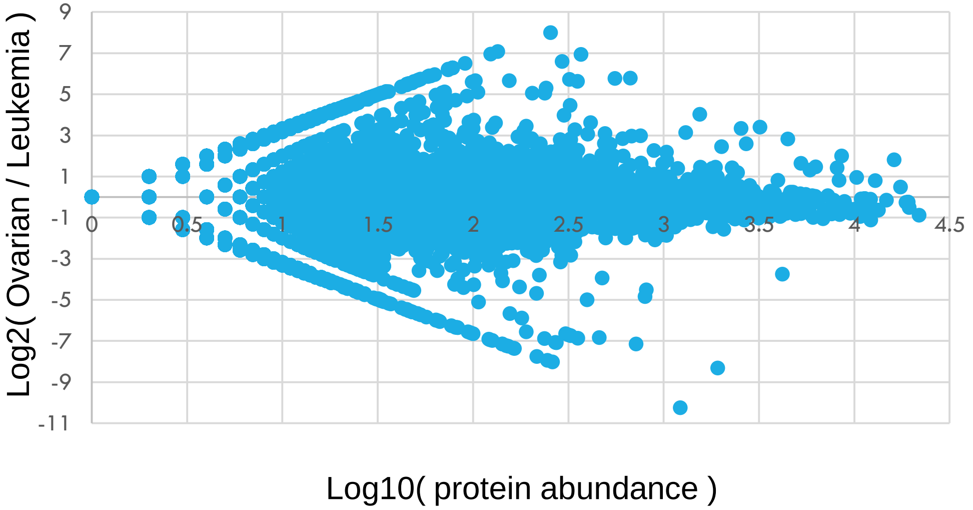
- 1,400,282 identified spectra
- 11,351 identified proteins
- Live results
  - G1: Leukemia
  - G2: Ovarian cancer



Differential protein expression



Protein ratios correlate with abundance



INTRODUCTION TO NCI 60 CELL LINES

Panel of cell lines established by the NIH National Cancer Institute (NCI) in the late 1980s to support the study of tumors

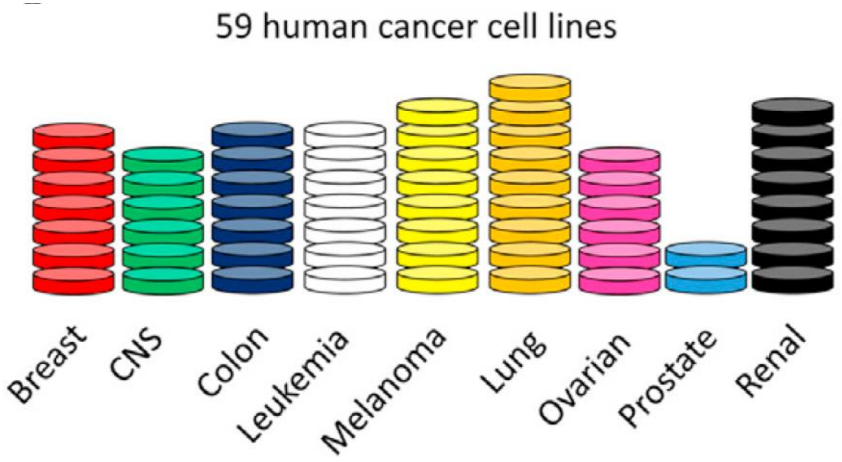
- Cell lines derived from 9 sources: Breast, CNS, Colon, Leukemia, Melanoma, Lung, Ovarian, Prostate and Renal cancers

Deep molecular characterization

- Genomics, gene expression (e.g., microarrays, transcriptomics, etc.), proteomics, metabolomics, etc.

Drug resistance and sensitivity

- Tested against >21,000 compounds



Shoemaker 2006, Nat Rev Cancer 6(10):813-23

NCI60 DRUG TREATMENTS

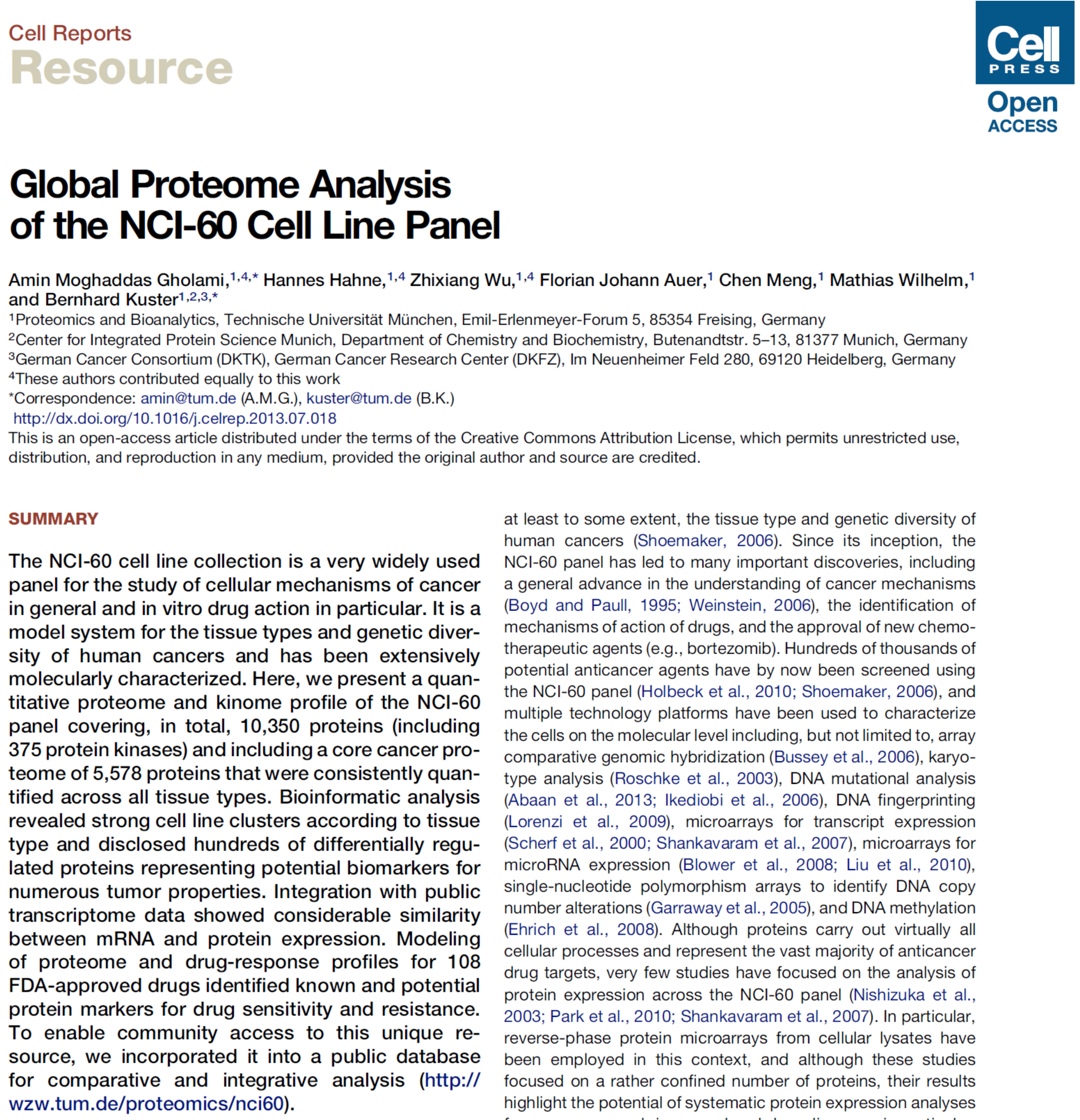
Compound activity has been extensively probed by treating NCI60 cell lines with clinical, pre-clinical and other compounds

- 187 FDA-approved
- 75 were in clinical trials
- 21,476 other compounds

Compounds are added to NCI60 cell cultures and observed phenotypes passing QC are added to a community collection of curated experiments

- NIH Developmental Therapeutics Program (DTP) – official NIH program managing the resource and distributing raw data for all activity probes
- CellMiner – aggregator site with processed data and transformed views designed to facilitate reutilization of NCI60 data

Shoemaker 2006, Nat Rev Cancer 6(10):813-23; Reinhold 2012, Cancer Res. 72(14):3499-511.



Cell Reports
Resource

Global Proteome Analysis of the NCI-60 Cell Line Panel

Amin Moghaddas Gholami,[1,4,*] Hannes Hahne,[1,4] Zhixiang Wu,[1,4] Florian Johann Auer,[1] Chen Meng,[1] Mathias Wilhelm,[1] and Bernhard Kuster[1,2,3,*]

[1] Proteomics and Bioanalytics, Technische Universität München, Emil-Erlenmeyer-Forum 5, 85354 Freising, Germany
[2] Center for Integrated Protein Science Munich, Department of Chemistry and Biochemistry, Butenandtstr. 5–13, 81377 Munich, Germany
[3] German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[4] These authors contributed equally to this work
*Correspondence: amin@tum.de (A.M.G.), kuster@tum.de (B.K.)

http://dx.doi.org/10.1016/j.celrep.2013.07.018

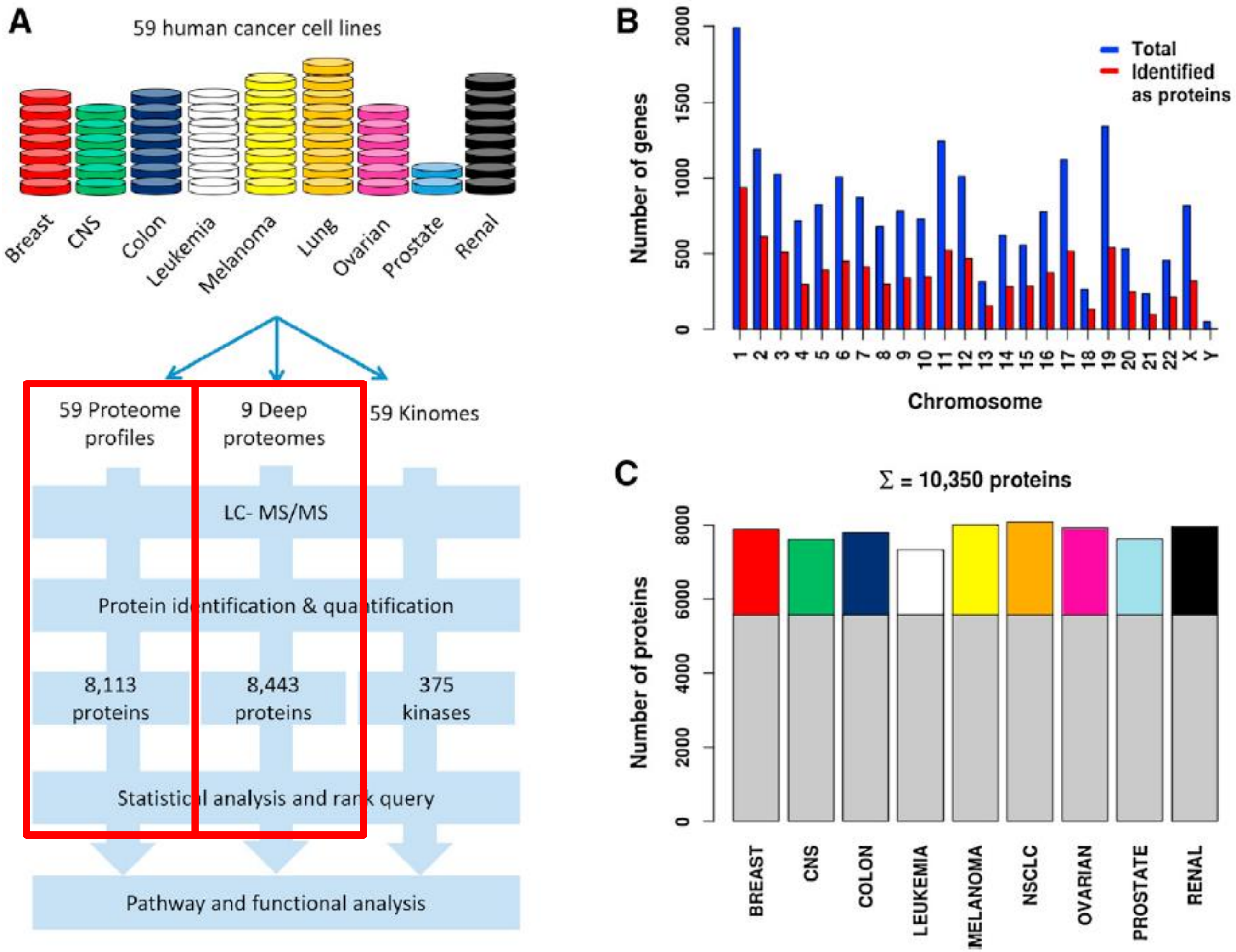
SUMMARY

The NCI-60 cell line collection is a very widely used panel for the study of cellular mechanisms of cancer in general and in vitro drug action in particular. It is a model system for the tissue types and genetic diversity of human cancers and has been extensively molecularly characterized. Here, we present a quantitative proteome and kinome profile of the NCI-60 panel covering, in total, 10,350 proteins (including 375 protein kinases) and including a core cancer proteome of 5,578 proteins that were consistently quantified across all tissue types. Bioinformatic analysis revealed strong cell line clusters according to tissue type and disclosed hundreds of differentially regulated proteins representing potential biomarkers for numerous tumor properties. Integration with public transcriptome data showed considerable similarity between mRNA and protein expression. Modeling of proteome and drug-response profiles for 108 FDA-approved drugs identified known and potential protein markers for drug sensitivity and resistance. To enable community access to this unique resource, we incorporated it into a public database for comparative and integrative analysis (http://wzw.tum.de/proteomics/nci60).

at least to some extent, the tissue type and genetic diversity of human cancers (Shoemaker, 2006). Since its inception, the NCI-60 panel has led to many important discoveries, including a general advance in the understanding of cancer mechanisms (Boyd and Paull, 1995; Weinstein, 2006), the identification of mechanisms of action of drugs, and the approval of new chemotherapeutic agents (e.g., bortezomib). Hundreds of thousands of potential anticancer agents have by now been screened using the NCI-60 panel (Holbeck et al., 2010; Shoemaker, 2006), and multiple technology platforms have been used to characterize the cells on the molecular level including, but not limited to, array comparative genomic hybridization (Bussey et al., 2006), karyotype analysis (Roschke et al., 2003), DNA mutational analysis (Abaan et al., 2013; Ikediobi et al., 2006), DNA fingerprinting (Lorenzi et al., 2009), microarrays for transcript expression (Scherf et al., 2000; Shankavaram et al., 2007), microarrays for microRNA expression (Blower et al., 2008; Liu et al., 2010), single-nucleotide polymorphism arrays to identify DNA copy number alterations (Garraway et al., 2005), and DNA methylation (Ehrich et al., 2008). Although proteins carry out virtually all cellular processes and represent the vast majority of anticancer drug targets, very few studies have focused on the analysis of protein expression across the NCI-60 panel (Nishizuka et al., 2003; Park et al., 2010; Shankavaram et al., 2007). In particular, reverse-phase protein microarrays from cellular lysates have been employed in this context, and although these studies focused on a rather confined number of proteins, their results highlight the potential of systematic protein expression analyses

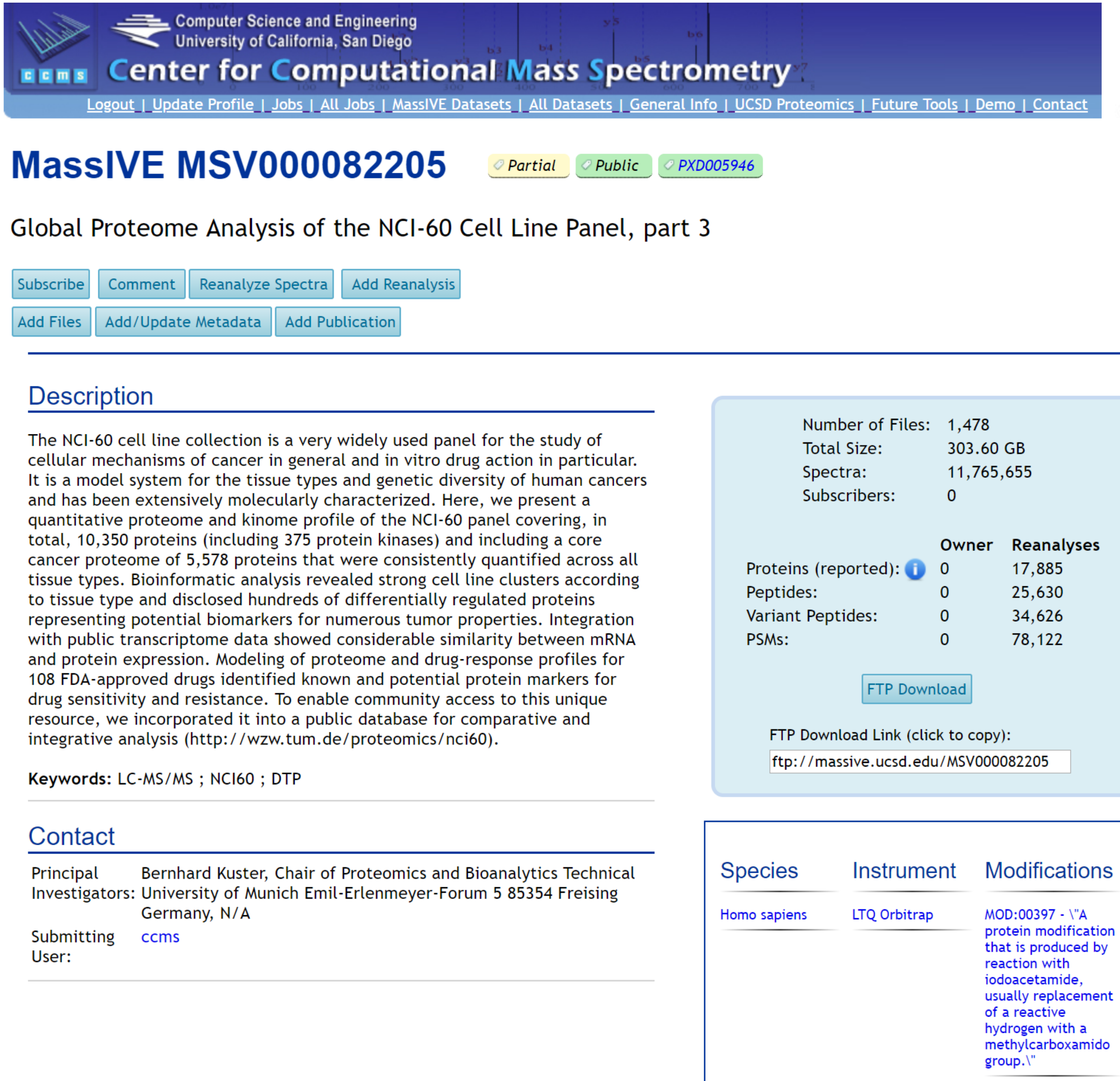
Gholami 2013, Cell Rep. 4(3):609-20

PROTEOME PROFILING OF NCI60 CELL LINES



Gholami 2013, Cell Rep. 4(3):609-20

NCI-60 DATASETS



Computer Science and Engineering
University of California, San Diego
Center for Computational Mass Spectrometry

Logout | Update Profile | Jobs | All Jobs | MassIVE Datasets | All Datasets | General Info | UCSD Proteomics | Future Tools | Demo | Contact

MassIVE MSV000082205

Partial Public PXD005946

Global Proteome Analysis of the NCI-60 Cell Line Panel, part 3

Description

The NCI-60 cell line collection is a very widely used panel for the study of cellular mechanisms of cancer in general and in vitro drug action in particular. It is a model system for the tissue types and genetic diversity of human cancers and has been extensively molecularly characterized. Here, we present a quantitative proteome and kinome profile of the NCI-60 panel covering, in total, 10,350 proteins (including 375 protein kinases) and including a core cancer proteome of 5,578 proteins that were consistently quantified across all tissue types. Bioinformatic analysis revealed strong cell line clusters according to tissue type and disclosed hundreds of differentially regulated proteins representing potential biomarkers for numerous tumor properties. Integration with public transcriptome data showed considerable similarity between mRNA and protein expression. Modeling of proteome and drug-response profiles for 108 FDA-approved drugs identified known and potential protein markers for drug sensitivity and resistance. To enable community access to this unique resource, we incorporated it into a public database for comparative and integrative analysis (http://wzw.tum.de/proteomics/nci60).

Keywords: LC-MS/MS ; NCI60 ; DTP

Contact

Principal Investigators:	Bernhard Kuster, Chair of Proteomics and Bioanalytics Technical University of Munich Emil-Erlenmeyer-Forum 5 85354 Freising Germany, N/A
Submitting User:	ccms

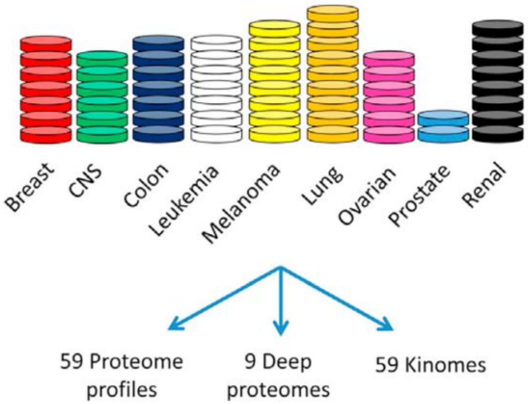
Number of Files:	1,478
Total Size:	303.60 GB
Spectra:	11,765,655
Subscribers:	0

	Owner	Reanalyses
Proteins (reported):	0	17,885
Peptides:	0	25,630
Variant Peptides:	0	34,626
PSMs:	0	78,122

FTP Download Link (click to copy):

ftp://massive.ucsd.edu/MSV000082205

Species	Instrument	Modifications
Homo sapiens	LTQ Orbitrap	MOD:00397 - \"A protein modification that is produced by reaction with iodoacetamide, usually replacement of a reactive hydrogen with a methylcarboxamido group.\"



Deep analysis
(MSV000082204)

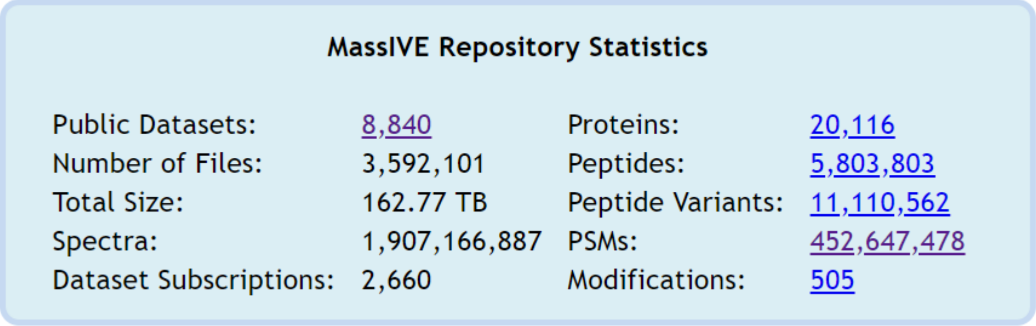
Profile analysis
(MSV000082205)

Kinome analysis
(MSV000082203)

GLOBAL MASS SPECTROMETRY BIG DATA

MassIVE

Mass Spectrometry
Interactive **V**irtual **E**nvironment



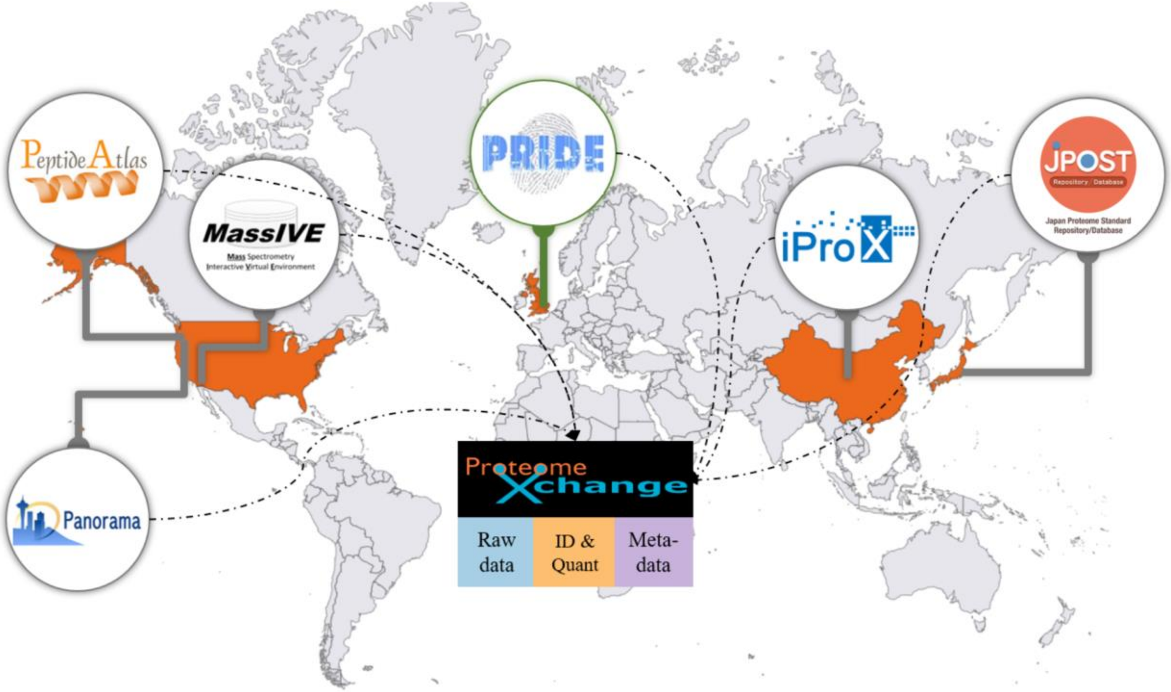
MassIVE Repository Statistics

Public Datasets:	8,840	Proteins:	20,116
Number of Files:	3,592,101	Peptides:	5,803,803
Total Size:	162.77 TB	Peptide Variants:	11,110,562
Spectra:	1,907,166,887	PSMs:	452,647,478
Dataset Subscriptions:	2,660	Modifications:	505

Full member of the

ProteomeXchange

consortium



BIOMOLECULAR SYSTEMS BEYOND JUST "BIG DATA"

Big Data

MassIVE

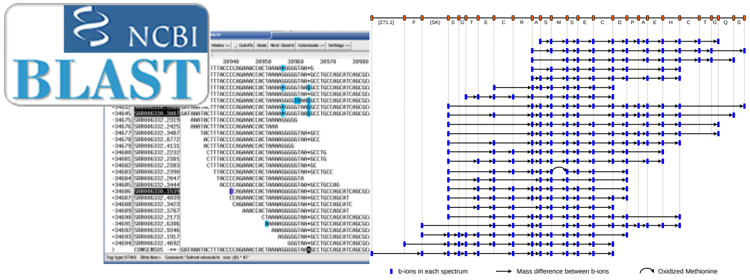
proteome CHANGE

Mass Spectrometry
Interactive Virtual Environment

Thousands of datasets, hundreds of terabytes

http://massive.ucsd.edu

Big Algorithms

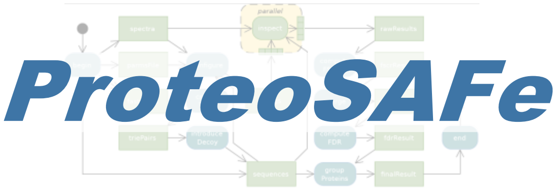


Designed to build on rather than just 'tolerate' big data

http://proteomics.ucsd.edu/software

Big Compute

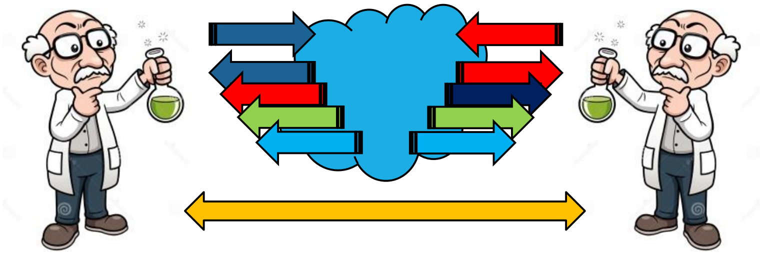
Proteomics Scalable, Accessible and Flexible environment



50+ data analysis workflows scalable to thousands of cores

http://proteomics.ucsd.edu/ProteoSAFe

Big Community



Empower and enable community-wide sharing of knowledge

http://gnps.ucsd.edu