Approaches to uncovering the 'dark' proteome

Tim Griffin
BMBB retreat
September 30, 2017
tgriffin@umn.edu

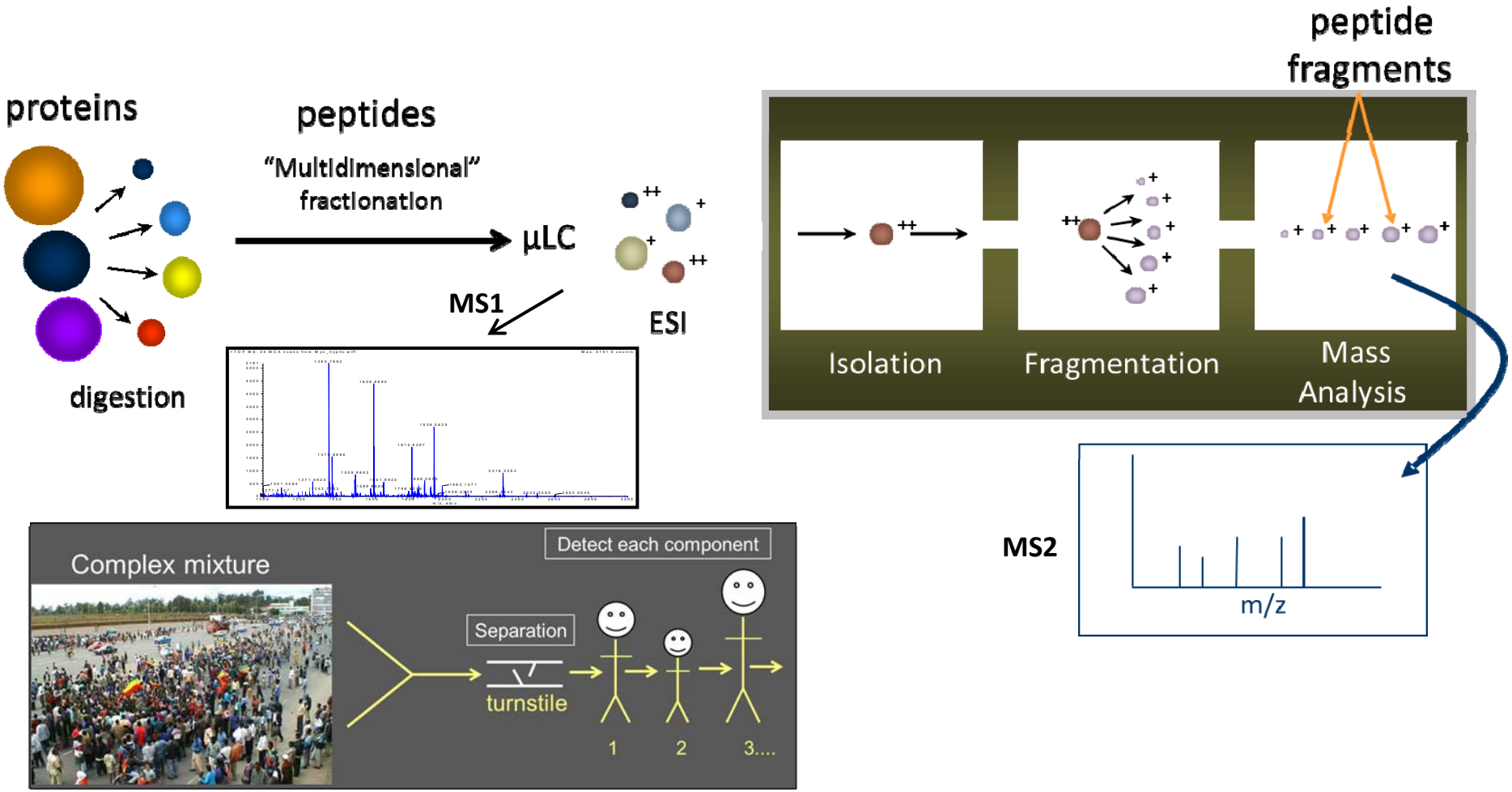
Outline

- Integrative genomic-proteomic for mapping novel protein sequence variants (aka proteogenomics)

- Increasing sequence coverage of proteins (aka finding tryptic peptides we often miss in MS-based proteomics)

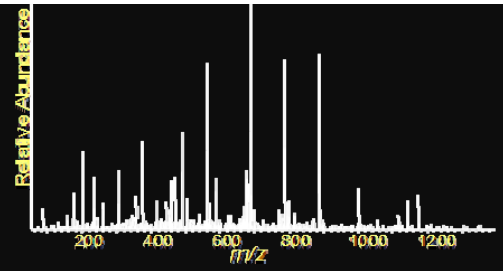
"Bottom-up" MS-based proteomics: driving technology

Peptide fractionation coupled to tandem mass spectrometry (MS/MS)



Using the data: from sticks on graph to protein identification

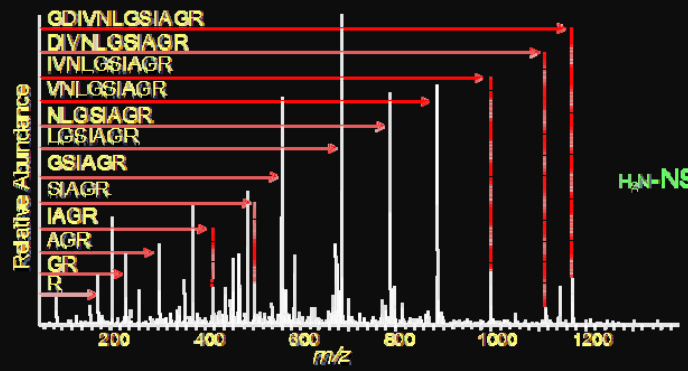
Raw MS/MS spectrum



Protein sequence and/or DNA sequence database search

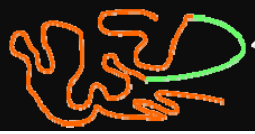
Direct identification of 1000s proteins from complex mixtures

Peptide sequence match

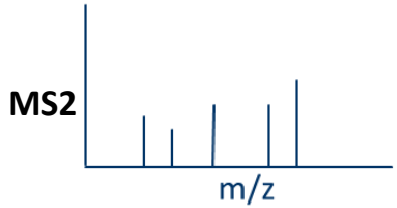


H_2N-NSGDIVNLGSIAGR-COOH

Protein identification



Inferring protein identity from peptide sequence matches

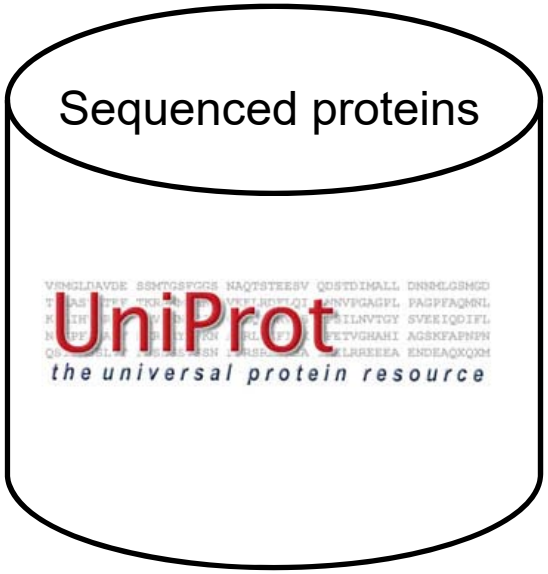


CAQCHTVEK

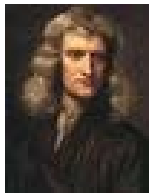
Cytochrome C

$_{NH_2}$GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGL
FGRKTGQAPGFTYTDANKNKGITWKEETLMEYLENPK
KYIPGTKMIFAGIKKKTEREDLIAYLKKATNE$_{COOH}$

Illuminating new protein sequence variants

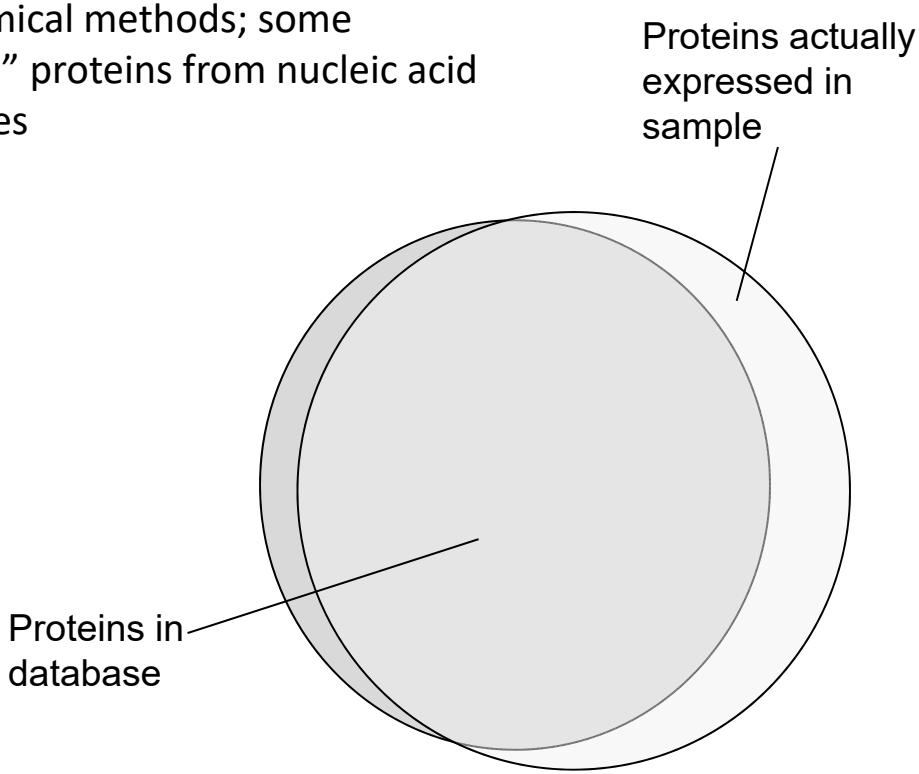


- Generally includes "reference" protein sequences or single proteins isolated via biochemical purification and chemical methods; some "inferred" proteins from nucleic acid sequences

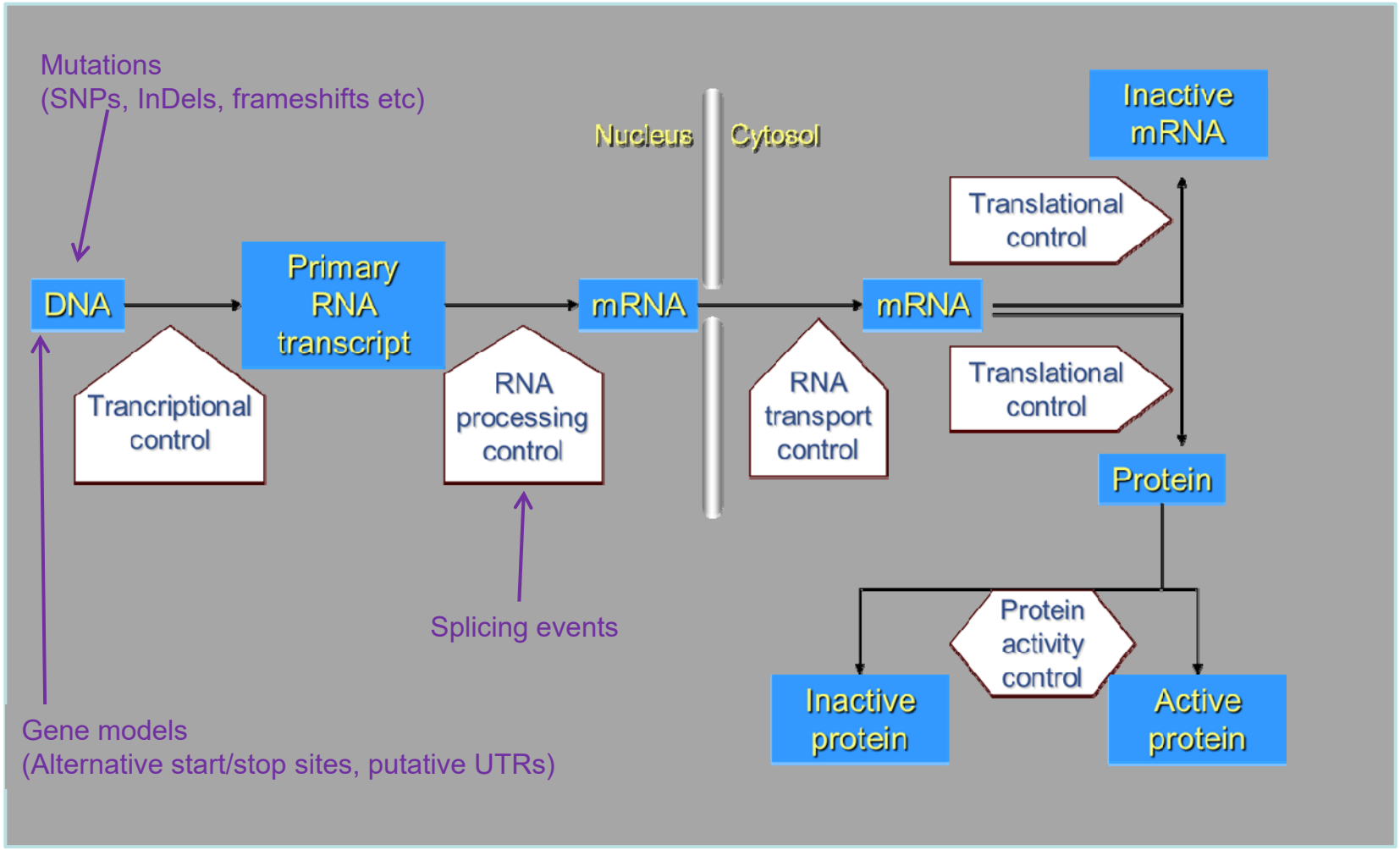


"What we know is a drop, what we don't know is an ocean."

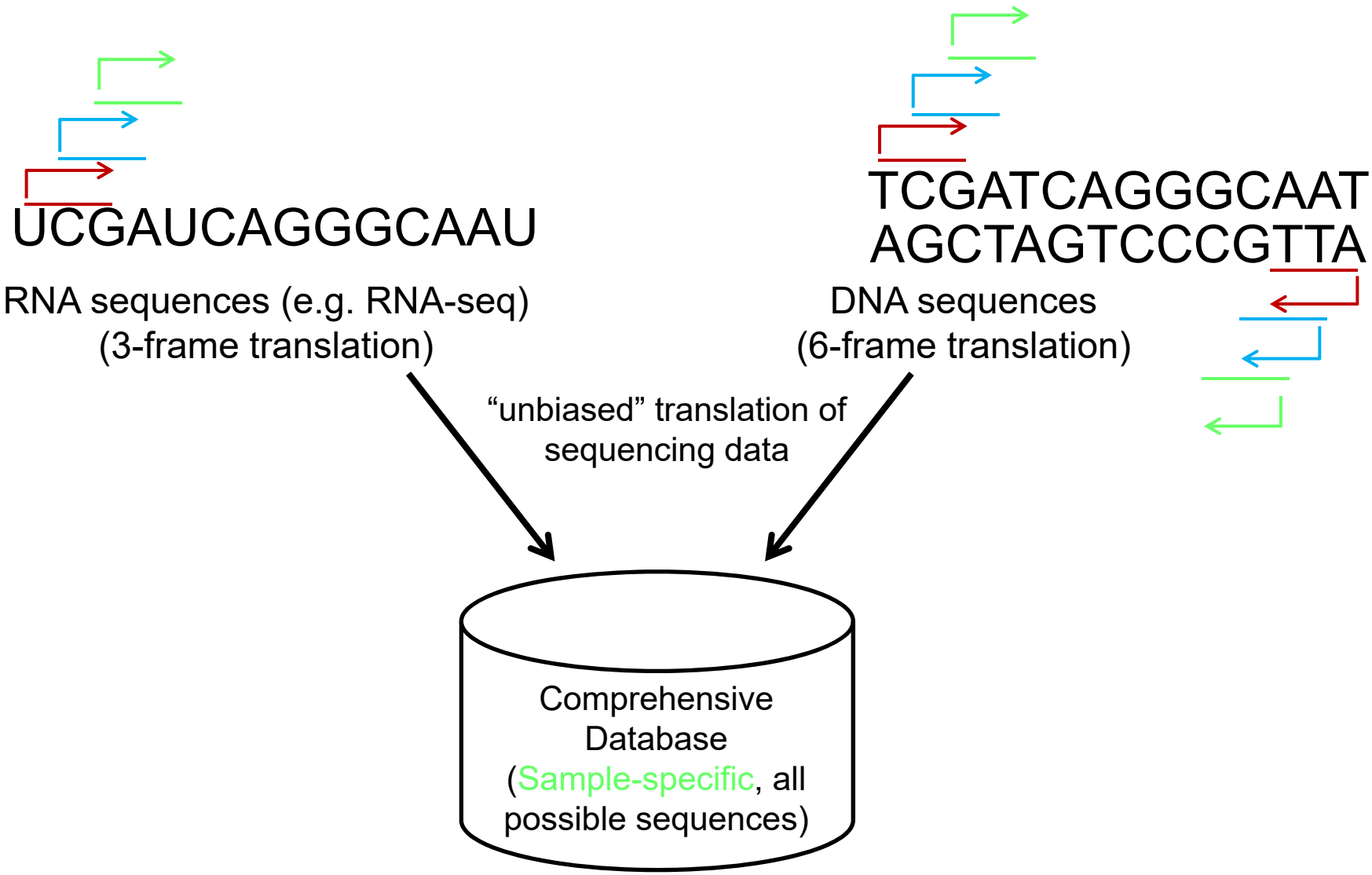
-- Sir Isaac Newton



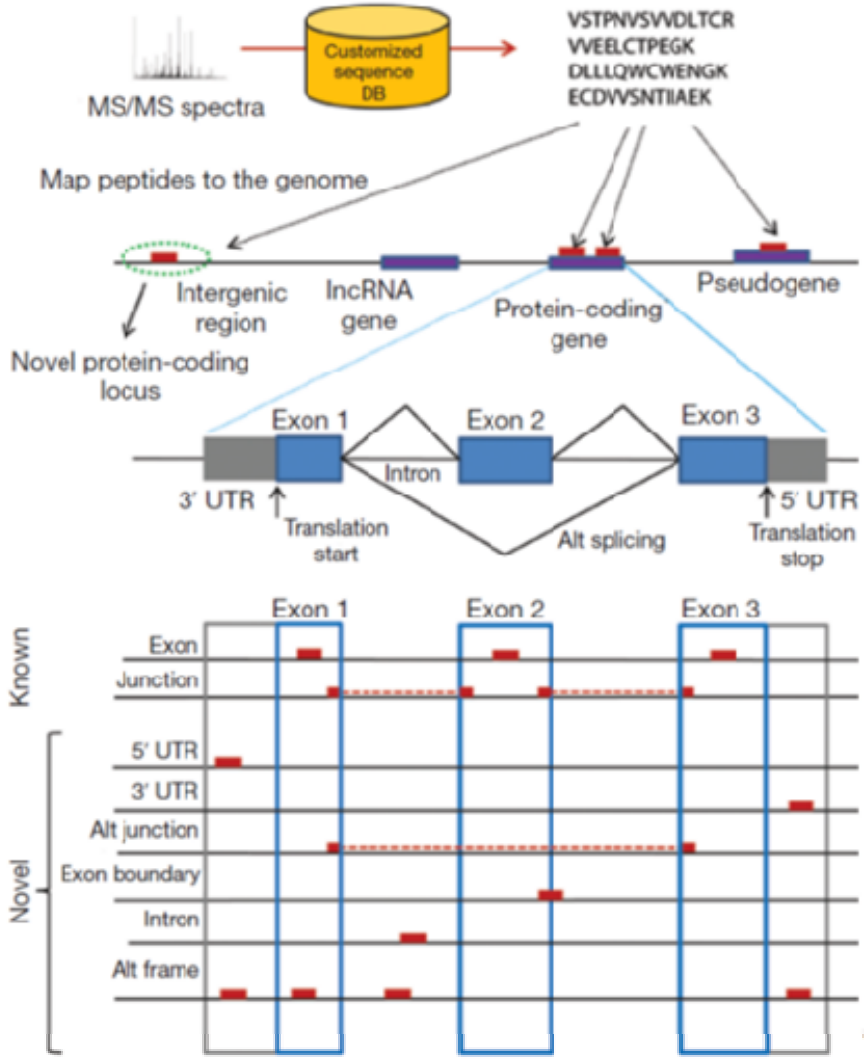
Sources of “novel” protein sequence variants?



Creating a more personalized protein sequence database



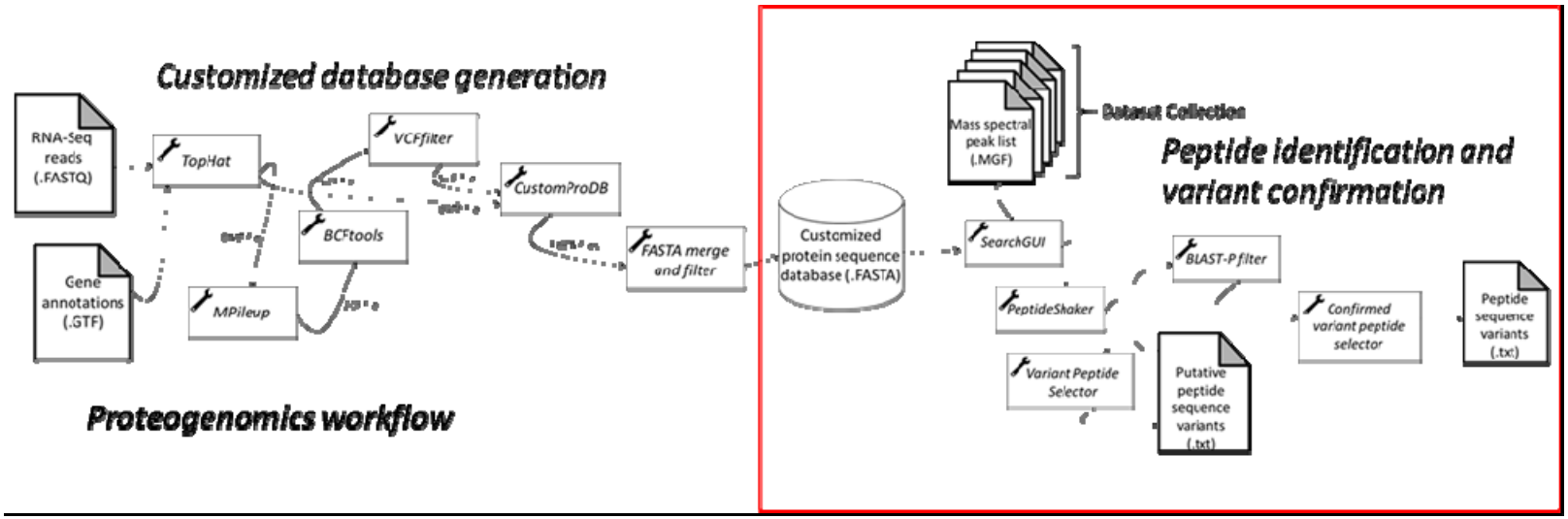
Proteogenomics: a multi-omics approach



- ✓ Genome annotation
- ✓ Gene expression regulation
- ✓ Protein variants in disease
- ✓ Functional outcomes of genome mutation

VOL.11 NO.11 | NOVEMBER 2014 | NATURE METHODS

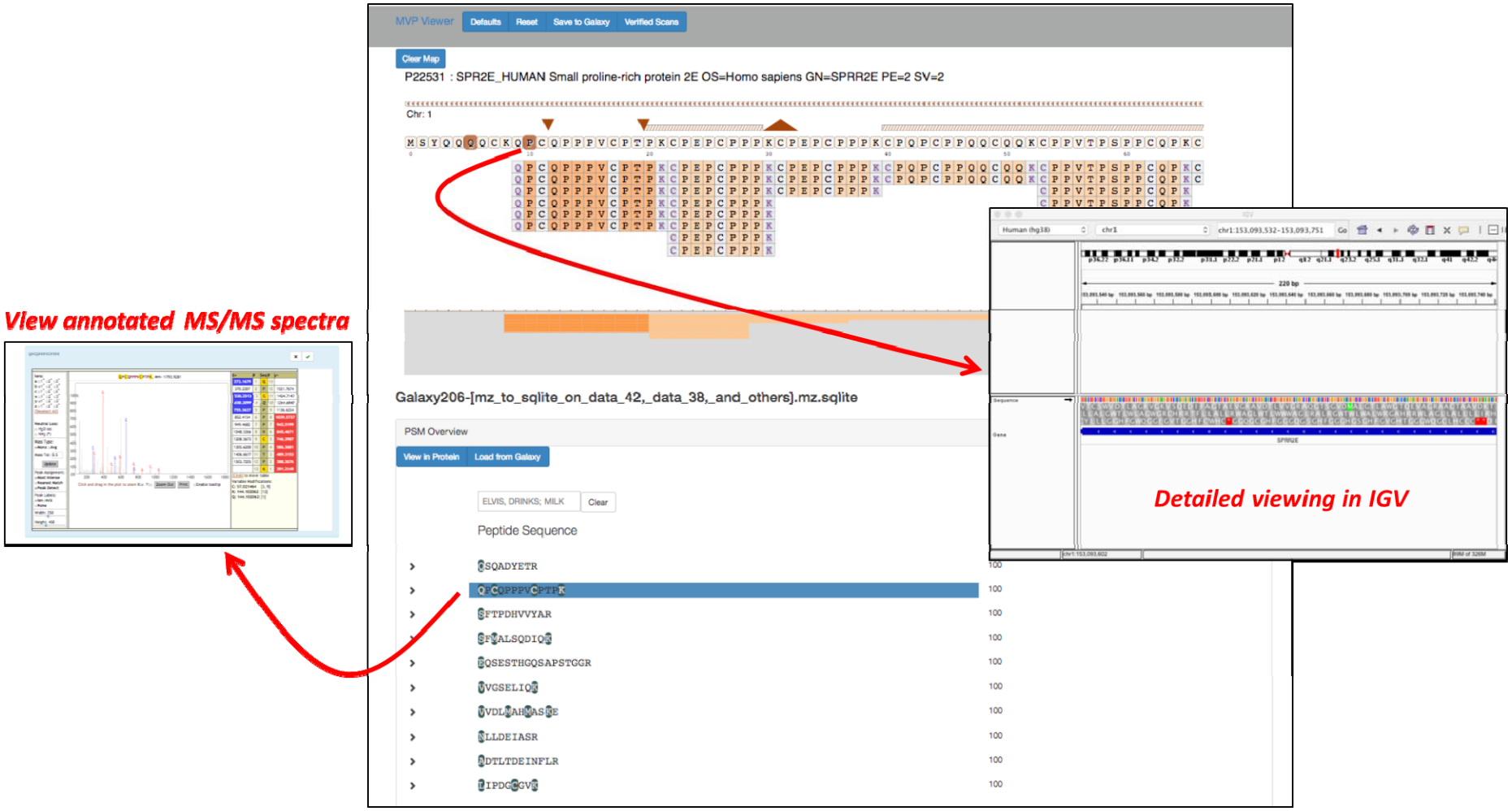
Workflow development: Galaxy-P



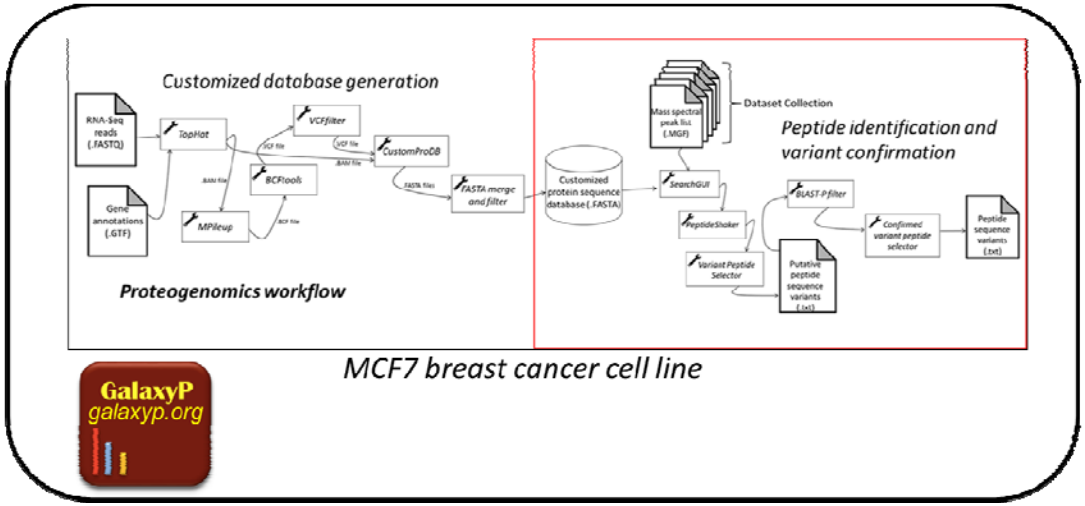
Cancer Research, in press

GalaxyP
galaxyp.org

Automated visualization and interpretation (in process)



Non-automated way: crowd-sourcing via first year grad students



Protein name	Protein function (brief description)	Variant
FoxK1	Transcriptional regulator	K261T
Histone H1.4	Histone H1 protein binds to linker DNA between nucleosomes	K160R
Nuclear Factor NF kB (kappa beta)	Transcription factor	V60A
DNA-Directed RNA polymerase I subunit RPA34	Catalyzes transcription of DNA into RNA	K261T
Fatty acid-binding protein, heart	Iintracellular transport of long-chain fatty acids	K64R

▪
▪
▪

Ongoing collaborative work

Proteogenomics

- Jeongsik Yong lab (cancer)
- Laurie Parker lab (cancer)
- Maneesh Barghava (lung disease)

⋮

Metaproteomics

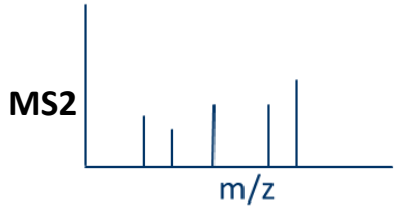
- Joel Rudney lab (periodontitis)
- Amy Skubitz (cancer)
- Terri Laguna (cystic fibrosis)

⋮

GalaxyP
galaxyp.org

More?

Illuminating the “dark” tryptic peptidome…



CAQCHTVEK

Cytochrome C

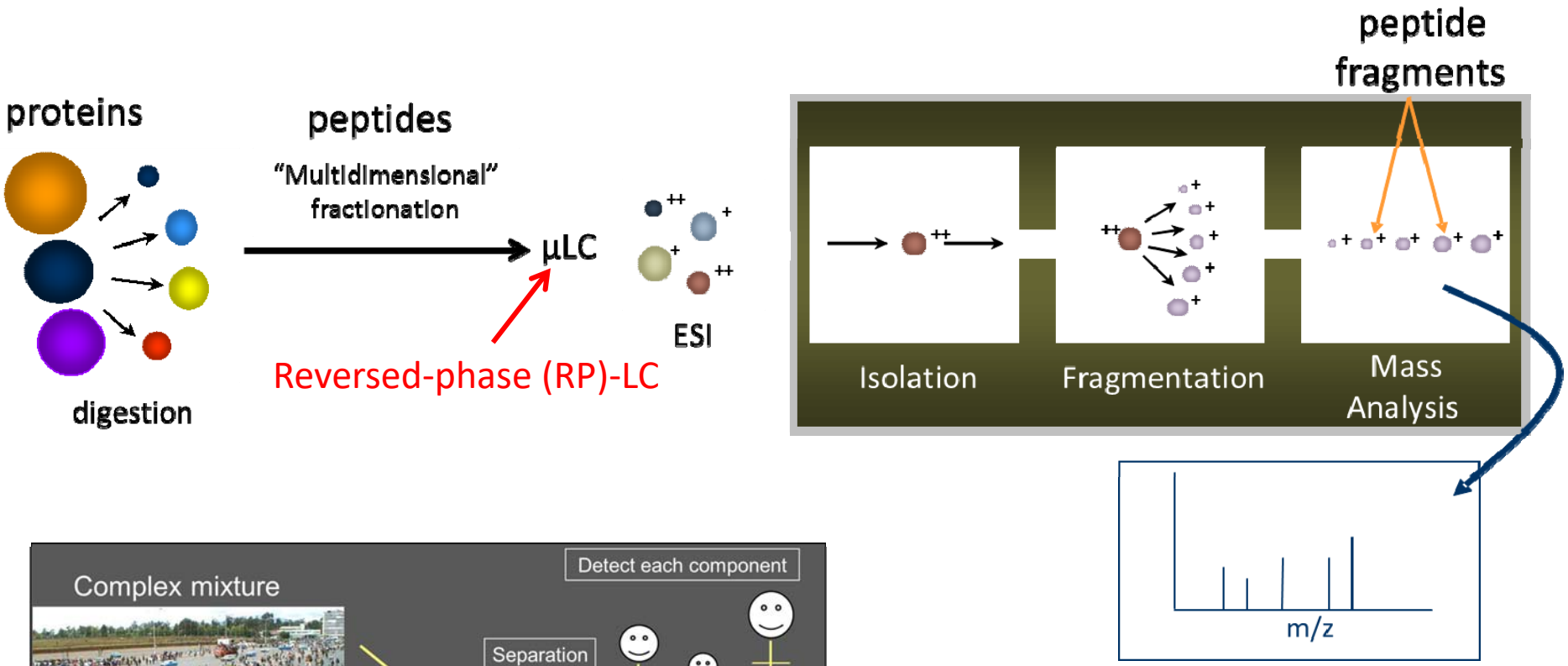
$_{NH_2}$GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGL
FGRKTGQAPGFTYTDANKNKGITWKEETLMEYLENPK
KYIPGTKMIFAGIKKKTEREDLIAYLKKATNE$_{COOH}$

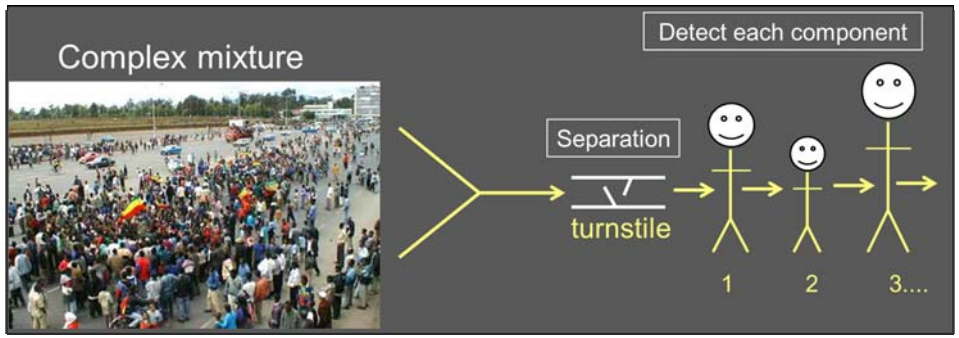
?

?

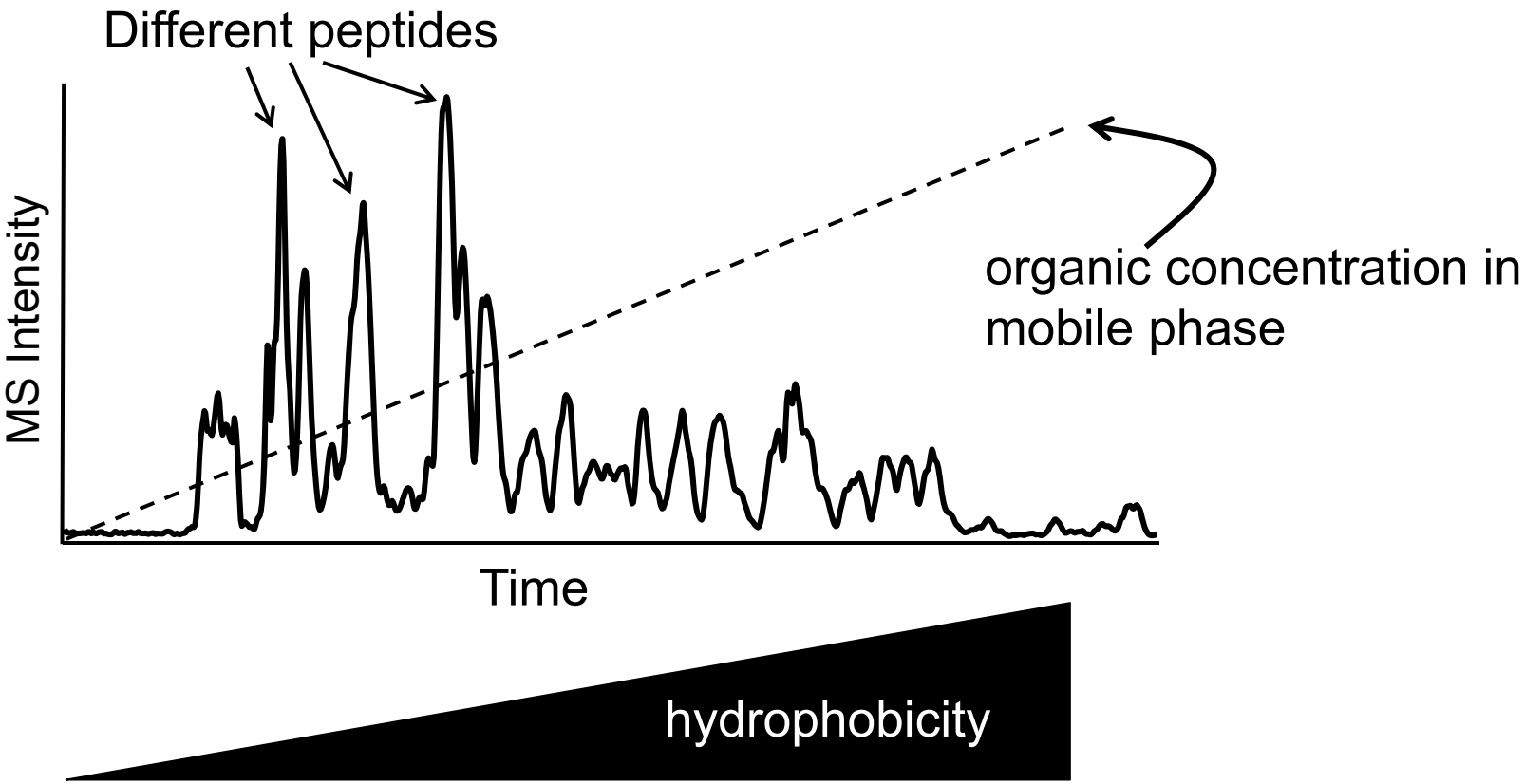
The underlying platform

Peptide fractionation coupled to tandem mass spectrometry (MS/MS)





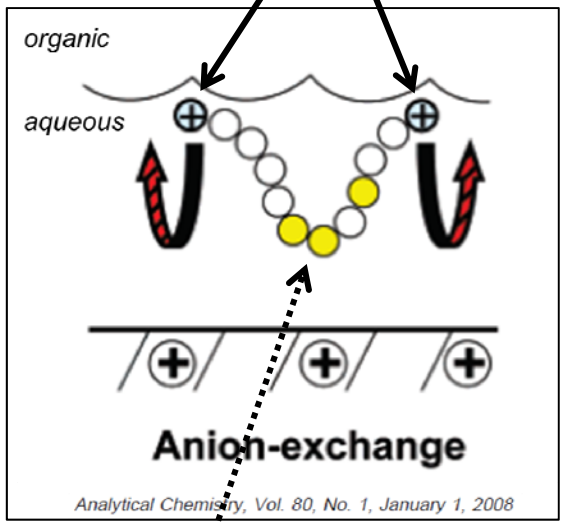
The standard: RP-LC-MS



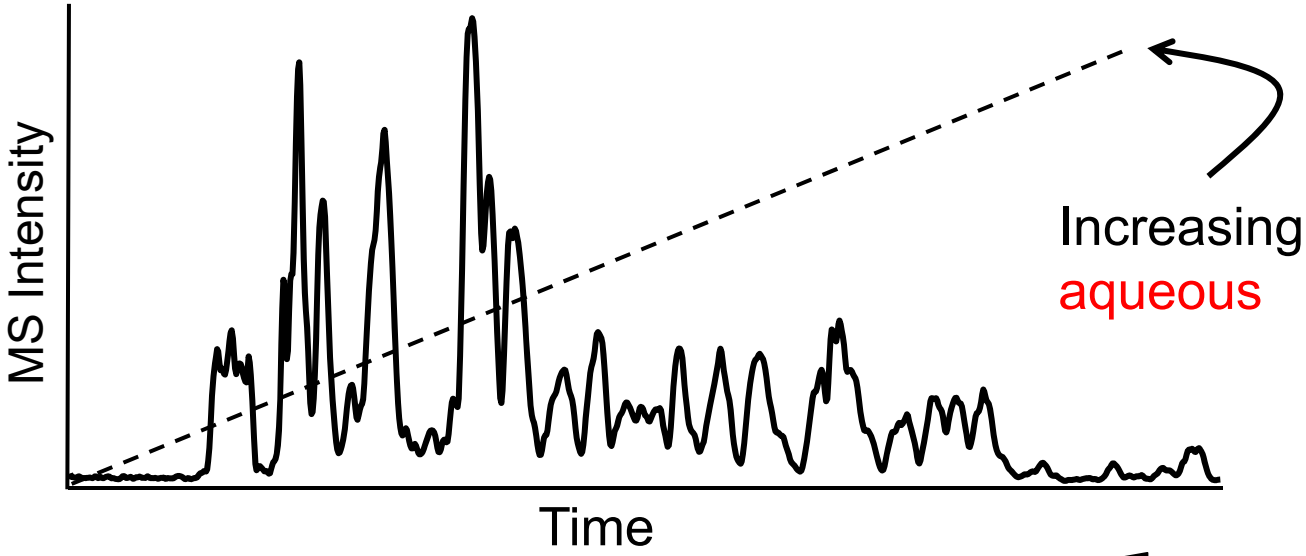
- Reverse-phase (RP) liquid chromatography (LC) peptide separation prior to online MS analysis

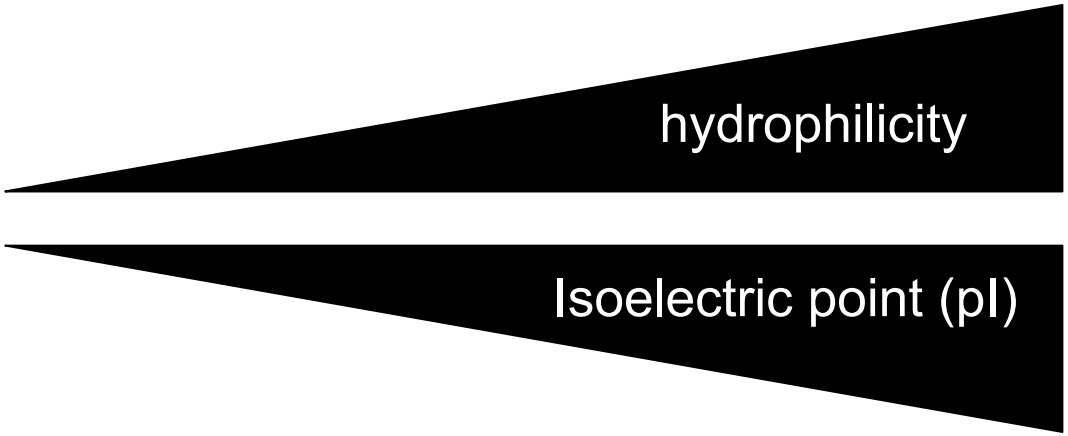
Electrostatic-Repulsion Hydrophilic Interaction Chromatography (**ERLIC**)

Basic sites (n-terminus, K, R, H)

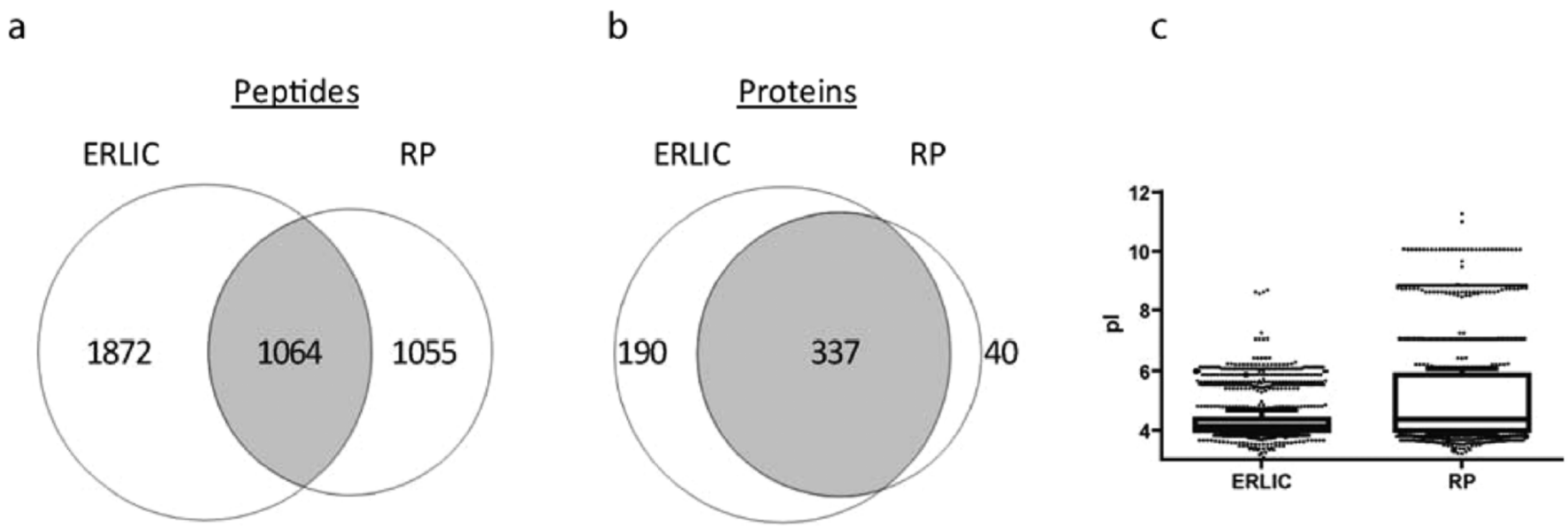


Polar and/or acidic sites





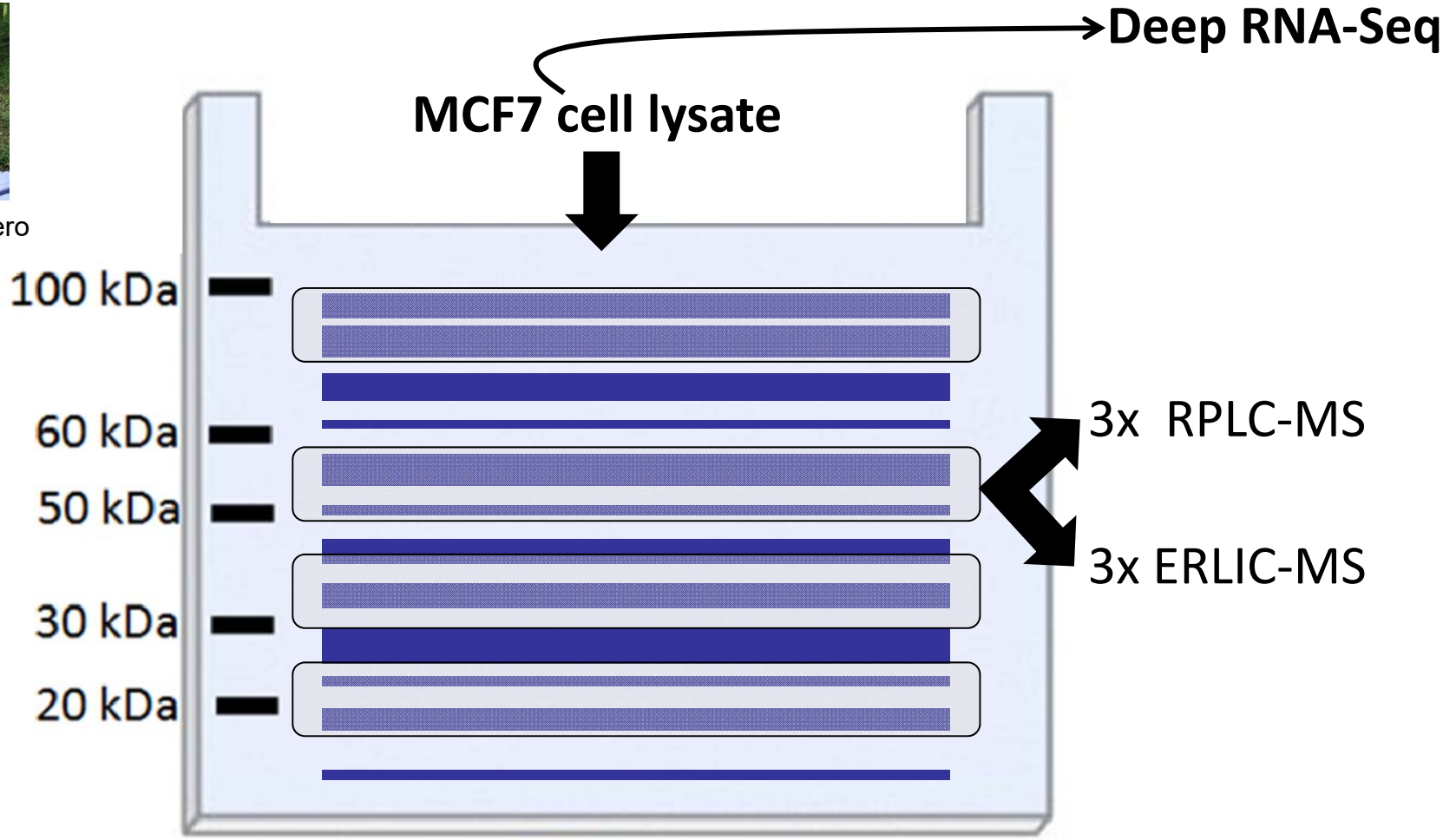
Initial results online ERLIC-MS: complementary identification of peptides



De Jong, E.P. and Griffin, T.J.
J. Proteome Res. 2012, 11, 5059–5064

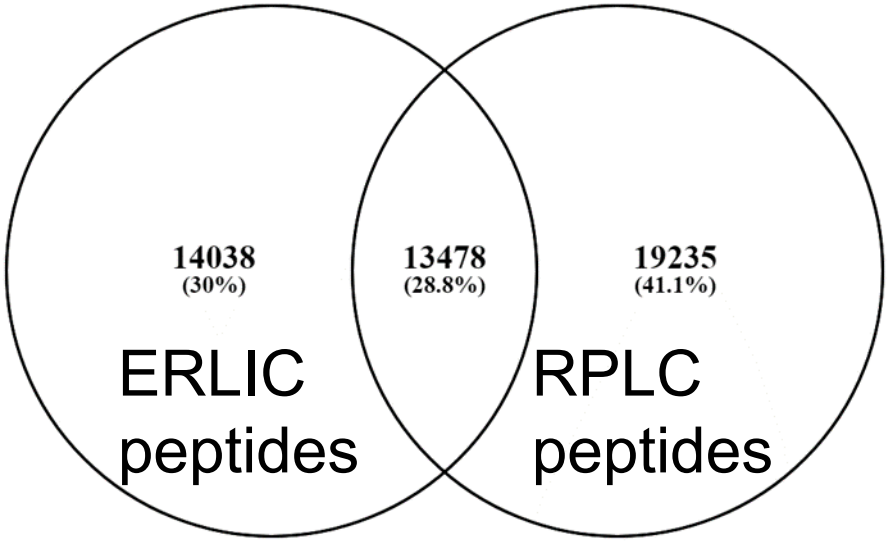
Testing ERLIC-MS on better instrumentation and new applications

Candace Guerrero



Confirming complementary identification of peptides

Orbitrap Fusion

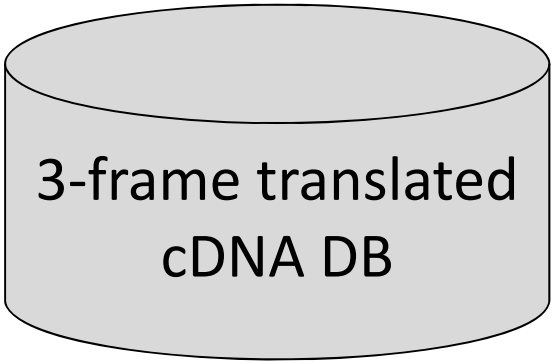


>sp|Q9Y6X8|ZHX2_HUMAN Zinc fingers and homeoboxes protein 2 OS=Homo sapiens GN=ZHX2 PE=1 SV=1 (837 AA)

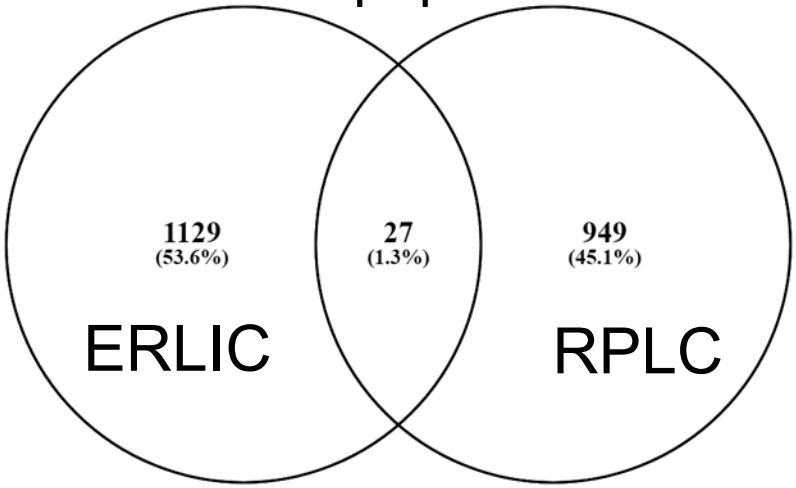
MASKRKSTTPCMVRTSQVVEQDVPEEVDRAKEKGIGTPQPDVAKDSWAAELENSSKENEVIE
VKSMGESQSKKLQGGYECKYCPYSTQNLNEFTEHVDMQHPNVILNPLYVCAECNFTTKKYDSL
SDHNSKFHPGEANFKLKLIKRNNQTVLEQSIETTNHVVSITTSGPGTGDSDSGISVSKTPIMKPG
KPKADAKKVPKKPEEITPENHVEGTARLVTDTAEILSRLGGVELLQDTLGHVMPSVQLPPNINLV
PKVPVPLNTTKYNSALDTNATMINSFNKFPYPTQAELSWLTAASKHPEEHIRIWFATQRLKHGI
SWSPEEVEEARKKMFNGTIQSVPPTITVLPAQLAPTKVTQPILQTALPCQILGQTSLVLTQVTSG
STTVSCSPITLAVAGVTNHGQKRPLVTPQAAPEPKRPHIAQVPEPPPKVANPPLTPASDRKKTK
EQIAHLKASFLQSQFPDDAEVYRLIEVTGLARSEIKKWFSDHRYRCQRGIVHITSESLAKDQLAIA
ASRHGRTYHAYPDFAPQKFKEKTQGQVKILEDSFLKSSFPTQAELDRLRVETKLSRREIDSWFSE
RRKLRDSMEQAVLDSMGSGKKGQDVGAPNGALSRLDQLSGAQLTSSLPSPSPAIAKSQEQVH
LLRSTFARTQWPTPQEYDQLAAKTGLVRTEIVRWFKENRCLLKTGTVKWMEQYQHQPMADD
HGYDAVARKATKPMAESPKNGGDVVPQYYKDPKKLCEEDLEKLVTRVKVGSEPAKDCLPAKPS
EATSDRSEGSSRDGQGSDENEESSVVDYVEVTVGEEDAISDRSDSWSQAAAEGVSELAESDSD
CVPAEAGQA

Blue: RP
Red: ERLIC
Green: Common

Expanding the scope of proteogenomics



Novel peptides



1	AEATESAMEK
2	CDVDIRKDLYTNTVLSGGTTMYPGIADR
3	CPEALFQPSFLGMESCGIHDTTFNSIMK
4	CTMEMYHKALSEALPGDNVGFNVK
5	DLYANTVLSGGTNMYPGITDR
6	DLYTSTVLSGGTTMYPGIADR
7	DSYVGDEAQTK
8	DSYVGEDAQSKR
9	EAFAVALKANNSMSK
10	EKLCYLVLDFEQEMATAASSSSLEK
11	FGEVIDCTIK
12	FRCPEALFQPSFLGMESCGIHDTTFNSIMK

⋮

Future

- Full analysis of variant classes identified between ERLIC and RPLC
- Application to PTMs: expanding phosphoproteome etc.
- Implement with multiple enzyme proteolysis
- Automation ERLIC-RPLC system