Approaches to uncovering the 'dark' proteome

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

- Proteins are digested into peptides.
- Peptides undergo "Multidimensional" fractionation.
- The fractionated peptides are then separated by microfluidic chromatography (μLC).
- The resulting peptides are subjected to Mass Analysis and fragmentation for peptide detection.

Diagram: Flowchart of the process from proteins to peptides to fragment detection.
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

Protein identification
Inferring protein identity from peptide sequence matches

Cytochrome C

NH$_2$GDVEKGGKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWKEETLMLEYLENPKKYIPGTMIFAGIKKKTEREDLIAYLKATNE$_{COOH}$
“What we know is a drop, what we don't know is an ocean.”

-- Sir Isaac Newton

- Generally includes “reference” protein sequences or single proteins isolated via biochemical purification and chemical methods; some “inferred” proteins from nucleic acid sequences.
Sources of “novel” protein sequence variants?

- Mutations (SNPs, InDels, frameshifts etc)
- Gene models (Alternative start/stop sites, putative UTRs)
- Splicing events
Creating a more personalized protein sequence database

UCGAUCAGGGGCAAU
RNA sequences (e.g. RNA-seq)
(3-frame translation)

TCGATCAGGGCAAT
AGCTAGTCCCGTGA
DNA sequences
(6-frame translation)

“unbiased” translation of sequencing data

Comprehensive Database
(Sample-specific, all possible sequences)
Proteogenomics: a multi-omics approach

- Genome annotation
- Gene expression regulation
- Protein variants in disease
- Functional outcomes of genome mutation
Workflow development: Galaxy-P

Customized database generation

Peptide identification and variant confirmation

Proteogenomics workflow

Cancer Research, in press
Automated visualization and interpretation (in process)

View annotated MS/MS spectra

Detailed viewing in IGV
Non-automated way: crowd-sourcing via first year grad students

Proteogenomics workflow

MCF7 breast cancer cell line

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

[Image of a table with protein names, functions, and variants]

University of Minnesota
Driven to Discover™
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...

Cytochrome C

\[ \text{NH}_2 \text{GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGF} \text{TYTDANKNKGITWKEETLMYELENPKKYIPGTMIFAGIKKKTEREDLIAYLK} \text{KATNECOOH} \]
The underlying platform

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

proteins

peptides

µLC

“Multidimensional” fractionation

Reversed-phase (RP)-LC

Isolation
Fragmentation
Mass Analysis

peptide fragments

Complex mixture

Separation

Detect each component

turnstile

1 2 3 ....
The standard: RP-LC-MS

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

Different peptides

organic concentration in mobile phase

hydrophobicity
Electrostatic-Repulsion Hydrophilic Interaction Chromatography (ERLIC)

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

Polar and/or acidic sites

Increasing aqueous

Anion-exchange

MS Intensity

Time

Increasing aqueous

hydrophilicity

Isoelectric point (pI)
De Jong, E.P. and Griffin, T.J.

*J. Proteome Res.* 2012, 11, 5059–5064
Testing ERLIC-MS on better instrumentation and new applications

MCF7 cell lysate

Deep RNA-Seq

3x RPLC-MS

3x ERLIC-MS

Candace Guerrero
Confirming complementary identification of peptides

**Orbitrap Fusion**

![Orbitrap Fusion Image]

14038 (30%)
13478 (28.8%)
19235 (41.1%)

ERLIC peptides
RPLC peptides

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

MASRKRSTTPCMVRSTQVVEQDVPEEVEDRAKEKGITGPQPDVAKDSWAELSENSSKENEVIE
VKSMGESOSKSQLQQYECYCPYSTQNLENEFTEHVDMMQHPNVLNPVLVAECNFTTKYDSL
SDHNSKFHPGEANFKLKLKRNQTVLEQSIETTTHVVISNTSGPGLTSDDSGISYVSKTPIMKPG
KPKADAKKVPKKPEELTPENHVEGETARLVTDTAEILSRLGGEVEQDTLGHVMPSVQLPPNINLY
PKVCPVPLNTKYNSALDTNAIMINSFNNKFPYPTQAEWSWLTASKHPHEERIRWFATQRLKHGI
SWSSPEEVEEARKKMFGNTIQSVPPPTITVPQAPLAPTKVTCPTILQPCICLQLQTSLVLTQCVTSG
STTVSCSPTLAVAGVNTNHGQQRPFLVTQPAAPEPKRPIAQPPEPPEPANPVPLTPSDFRKKTK
EQIAHLKASFLQSDQPDADLYRLMGTGQSLVTRSEIKWFSDFRHRCQGRVHTSESLEKQDLAI
ASYRHGRTYHAYPDFAPQIKFKEKTQGOVQKILEDSSLKSSFPQTQAEELDRLVETKLSREEIDSWFSE
RRKLRDSSMEQAVLDSMGSGKKGQDVGAAPNGALSRLDOLSGAQLTSSLPSPSPAAKQOEYH
LLRSTFARTQWTPQEPYQDQLAATGTLLRTYETRVRWFKENRCLKLTKGTVKWMEQYQHQPMAADD
HGYDARVARKTTPMAESPKNGGDVDPQYKDPKLCLEEDELEKLVTRVKVGESEPADCLPAPKS
EATSRSEGSRSRQGSDENESSSSSVYDYVETVGEEDASDRSDSWSQAAEGYVSELAEDSDDCVPAEAGQA

Blue: RP
Red: ERLIC
Green: Common
Expanding the scope of proteogenomics

3-frame translated cDNA DB

Novel peptides

ERLIC

- 1129 (53.6%)
- AEATESAMEK
- CDVDIRKDLYTNTVLSGGTMYPGIADR
- CPEALFQPSFLGEMESCGIHDFTFNSIMK
- CTMEMYHKALSEALPGDNVGFN

RPLC

- 27 (1.3%)
- DLYANTVLSGTNMYPG
- DLYTSTVLSGGTTMPGIADR
- DSYVGDEAQTK
- DSYVGEDAQSKR
- EAFVAULKANSMSK
- EKLCLYLDEFEQEMATAASSSLEK
- FGEVIDCTIK
- FRCPEALFQPSFLGEMESCGIHDFTFNSIMK

...
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