Adding Another Horse to the Carriage: Using ERLIC-MS with RP-MS to Help Carry the Load of Expanding Protein Sequence Coverage

C. R. Guerrero[1], S. Mehta[1], J. Johnson[2], M. Chambers[3], P. Jagtap[2] , T. J. Griffin[1,2]

[1]University of Minnesota, [2]Minnesota Supercomputing Institute, [3]Vanderbilt University

Introduction

- **"Multi-omics" applications such as, Proteogenomics, are implemented to bridge gaps between "omic" fields.**
- **Proteogenomics connects genomic/transcriptomic and proteomic data to elucidate novel sequence variants and annotate for genomes.**

- **These discoveries are dependent on high sequence coverage via shotgun proteomics.**
- **Reverse phase (RP) chromatography is a staple of shotgun proteomics. However, to help to increase protein sequence coverage a different type of online chromatography could be useful.**

- **Electrostatic hydrophilic repulsion chromatography (ERLIC) can be a complementary method for RP chromatography improving peptide IDs, ultimately improving protein sequence coverage.**

- **Coupling a new analytical method with an optimized informatics approach can improve proteogenomic results, such as single amino acid variants (SAVs).**

- **Proteogenomic workflows with in Galaxy-P platform has been created to investigate the relationship between genomics and proteomics.**

ERLIC

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

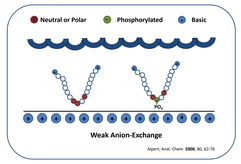
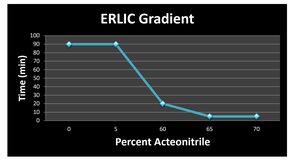


Figure 1. Peptide/stationary phase interaction with ERLIC chromatography

Two dimensions of separation
1. **Hydrophilic interaction**
2. **pI – isoelectric point**



Graph 1. ERLIC chromatography gradient with time vs percentage of Acteonitrile

Chromatography
- **PolySAX-LP Material**
- **High organic conditioning**
- **Elution with aqueous gradient**

Benefits
- **Retains more polar peptides**
- **Typically retains longer acidic peptide sequences**

Methods

MCF-7 Sample Preparation for Multiple Enzyme Digestion
- **200μg of MCF-7 protein lysate underwent an 18h in-solution digestion with either Trypsin/Lys-C, Trypsin, Elastase, or Pepsin**
- **Samples were ran individually by RPLC-MS/MS and ERLIC-MS/MS using data dependent mode**
- **All data was processed mudpit fashion with Peaks/Scaffold to assess protein sequence coverage increases**

MCF-7 Sample Preparation for Proteogenomics
- **50 μg of MCF-7 protein lysate and resolved via SDS-PAGE and 5 gel regions were selected for in-gel digestion with trypsin**
- **Each sample was analyzed by both RPLC-MS/MS and ERLIC-MS/MS using data dependent acquisition**
- **Raw data was analyzed using proteogenomic workflows within Galaxy-P. (Jagtap et al. 2014)**

Galaxy-P Framework

- **A web-based, community developed bioinformatics framework/platform/workbench**

- **Originally designed to address issues in genomic informatics including:**
  - **Software accessibility and usability**
  - **Analytical transparency**
  - **Reproducibility**
  - **Scalability**
  - **Share-ability**

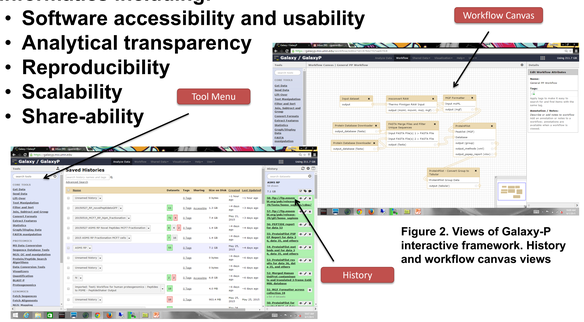


Figure 2. Views of Galaxy-P interactive framework. History and workflow canvas views

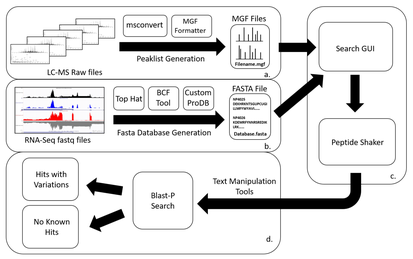
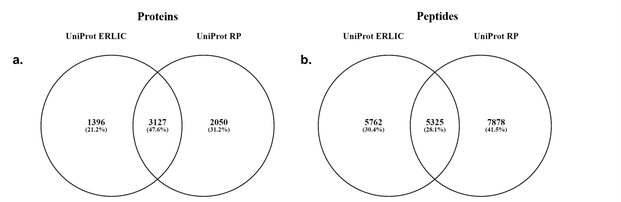
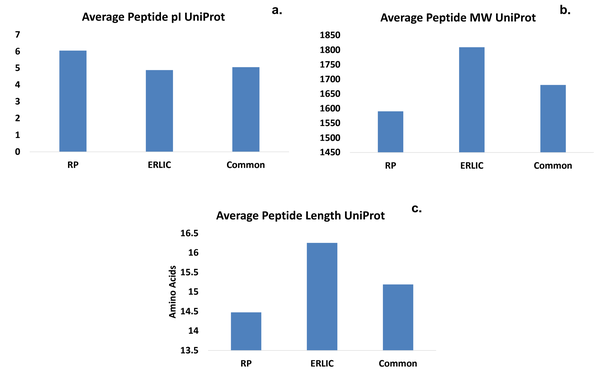


Diagram 1. Galaxy-P Proteogenomic Workflow. a. Peaklist Generation of raw files, b. RNA-seq protein database generation, c. protein search, d. data filtering

Results: ERLIC vs RP

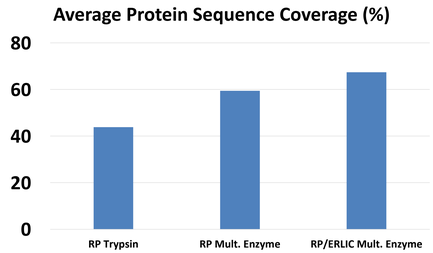


Graph 2. Five MCF-7 in-gel slices were digested with trypsin subsequently ran with RP- and ERLIC-MS/MS and analyzed using Search GUI/PeptideShaker with a human UniProt with cRAP database. Graph a. shows the protein identification associated with LC method while b. show peptide coverage. While roughly 48% of proteins are identified by both LC method, the lower overlap of peptides suggest that new peptides identified are associated with increasing protein sequence coverage of already existing proteins.



Graph 3. Five MCF-7 in-gel slices were digested with trypsin subsequently ran with RP- and ERLIC-MS/MS and analyzed using Search GUI/PeptideShaker with a human UniProt with cRAP database, a. shows average pI of peptides, b. illustrates average peptide MW, and c. shows average peptide length. These results fall in line with assumption ERLIC excels in retaining peptides with lower pI values, as well as, molecular weight which is indicative of longer peptides.

Results – Expanding Protein Sequence Coverage with Multiple Enzymes



Graph 4. MCF-7 lysate was digested with either Trypsin, Trypsin-Lys-C, Elastase, or Pepsin. All digested peptides mixtures were ran individually and processed as a whole in a PEAKS software using a human UniProt database. We show here the average protein sequence coverage of the top 100 proteins when using RP with trypsin alone, RP with multiple enzyme, an RP with ERLIC both using multiple enzymes. By combining RP and ERLIC-MS/MS data using multiple enzymes coverage was further improved by another 8 percent.

Results – Complementary ERLIC- and RPLC-MS/MS for Proteogenomic Assessment of MCF-7 cells

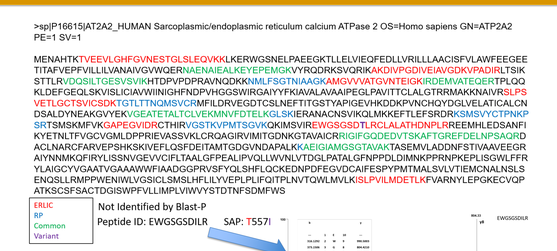
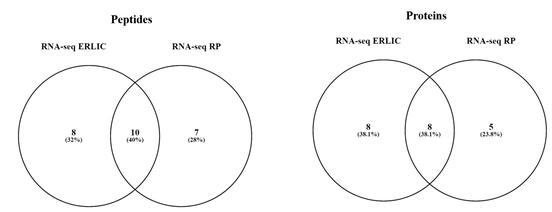


Figure 3. Protein sequence coverage of Sarcoplasmic/endoplasmic reticulum calcium ATPase 2. ERLIC peptides are in red, RP peptides are in blue, peptides identified by both are in green, and a single amino acid variant is represented in purple. EWGSGSDILR spectra shown as a match for single amino acid variant peptide

Novel RNA-seq Peptides Identified by each LC-MS Method		
RP	Common	ERLIC
AFEDDDITMVEGSVVPIR	EMFQSYDLDMMER	EAIQSPVSFLVLMNQIR
DAAAVGNHAAK	MVNSNLASVDELKEIDEVER	EWGSGSDILR
GFSDKLDFLEGDQKPLAQHK	MVNSNLASVDELKEIDEVERK	LLSSVFDEVYPALPSDVQTAIK
LLAEAGMPSPAWTK	STGVILANDANAER	NLEAVETLGSTSITCSDK
LLQSVGQAPESISEK	TDLNPDILQGGDDLDPNYVLSSR	NMTIGTAQADCAVLVAAGVGEFEAGISK
TPQEWAPMTAR	VADKIQLINNNLVK	SAIVMLINYQDDAELATHALPELTK
YSTGSDESFPMTTPSMR	YACEGPSMGGLPGASSEK	SDPVEFECINDK
	YMVPVVVVPEGSASNTMEQAILR	TPSAAYLWVGTGASEADKMGAQELLR
		MDGDKPSDEMKTDLNPDILQGGDDLDPNYVLSSR

Table 1. Shows the peptides where were identified as novel amino acid variants when using a RNA-seq derived protein database. ERLIC identified 8 variants that would have otherwise not by ID with RP-MS/MS alone.



Graph 3. Five MCF-7 in-gel slices were digested with trypsin subsequently ran with RP- and ERLIC-MS/MS and analyzed using Search GUI/Peptide Shaker with a RNA-seq derived database. Peptides were then filtered and sorted for best confidence and followed by screen of novel peptides via Blast-P. Peptides were either found via Blast-P with variation of a single amino acid variant or peptides were not identified at all.

Conclusions

- **Proteogenomic outcomes are dependent upon comprehensive protein sequence coverage for the detection of SAVs. ERLIC, as well as, multiple enzyme digestions offer an effective way to increase coverage**

- **ERLIC was able to identify eight variants that RP phase could not, offering a potential reason to include other efforts than RP-MS/MS alone**