

Characterization of sources of variability in LC-MS/MS analysis of protein complexes

W. Hayes McDonald¹, Trish K. Lankford², Gregory B. Hurst¹, Dale A. Pelletier², Linda J. Foote², Tse-Yuan S. Lu², Stephen J. Kennel²

Chemical Sciences Division¹, Life Sciences Division²,
Oak Ridge National Laboratory

OVERVIEW

- The goal of this effort was to begin to characterize some of the sources of variability in the LC-MS/MS analysis of protein complexes.

INTRODUCTION

- Mass spectrometry has long been used to identify and characterize proteins from isolated protein complexes. With the advent of new epitope tagging and molecular biology tools, it has become possible to undertake such efforts on a genome scale [1-3]. Through the use of dual-affinity tagging strategies such as the original tandem affinity purification (TAP) tag, the results of such purifications can be quite "clean" and reproducible [4]. Coupling these techniques to bottom up or shotgun proteomic analysis has proven to be incredibly powerful not only for the exquisitely sensitive identification of protein constituents, but also for characterization of their post-translational modifications (e.g. [5]). However, the challenge and utility of characterizing protein complexes is not just limited to well-studied model systems. The Center for Molecular and Cellular Systems (CMCS) (<http://www.ornl.gov/sci/Genomestolife/>) is a pilot project sponsored by Department of Energy Genomics:GTL program in order to develop the necessary infrastructure to study protein complexes from diverse sets of microorganisms. Part of this project involves the tagging, purification, and characterization of proteins expressed in *Rhodospseudomonas palustris*, a purple nonsulfur anoxygenic phototrophic bacterium that is ubiquitous in soil and water samples and noted for its extensive metabolic diversity. Since the final challenge is to bring together a high quality, well-validated data set, it is essential to begin to characterize the sources of variability in the analysis.

EXPERIMENTS

- Samples were isolated, prepared, and analyzed through the normal protocols of the CMCS "endogenous isolation" pipeline. Briefly, 1.5 liters of *R. palustris* cells expressing 6HIS/V5 dual tagged versions of either RPA2164 (GroEL1) or RPA3226 (rpoA) were grown, harvested, and lysed in duplicate. Proteins were isolated using sequential purifications first on nickel-NTA resin and then using conjugated anti-V5 epitope antibodies. A 20 percent aliquot was analyzed via Protein LabChip on an Agilent 2100 Bioanalyzer. This allowed measurements of both total and individual protein amounts (~2.5 ug total for the RPA2164 samples and ~13.5 ug total for the RPA3226). The remainder was denatured, digested with trypsin, acidified, and diluted for analysis. LC-MS/MS analysis was performed using a self-pack 16 cm reversed phase column (Jupiter C18, 5 um - Phenomenex) coupled directly to a Thermo LCQDecaXP nanospray source. Autosampling and HPLC separations used the LCPacking Ultimate-Famos-Switchos combination with a cycle time of approximately 3 hours per sample.

- For the purpose of this study, we analyzed in triplicate a 1/10 and 1/100 dilution from each of the two biological replicates of the two tagged strains - 24 samples in total. Sample order was randomized and a full blank injection analysis was included between each sample for a total analysis time of approximately 6 days. Resulting MS/MS spectra were searched using SEQUEST in "fully tryptic" mode against a predicted *R. palustris* protein database to which common contaminant proteins had been appended. Identifications were filtered and organized using DTASelect and Contrast [6]. Based on prior work using a similar instrument configuration on "known" protein standards we chose the following Xcorr cutoffs (+1 1.899 +2 2.029 +3 2.714) [7].

RESULTS AND DISCUSSIONS

Because of its known members, we focused our strategies for visualizing the consistency of results to the *R. palustris* RNA polymerase complex (tagged component rpoA). Graphs are shown for each of the 12 samples and an "average" graph for each triplicate.

Results: Protein Purification Consistency

LabChip analysis of the individual isolates indicate that they were consistent in total yield and in the relative complexity of the prominent proteins. However, based on the peak detection and quantitation data, the relative amounts of individual components appears to vary between the two biological replicates.

rpoA-6xHIS/V5 isolate 1

Size [kDa]	Rel. Conc. [ug/ml]	Rel. molarity [uM]	% Total
5	0	0	0
9.1	0	0	0
10.8	0	0	0
16.1	9.5	0.59	2
24.6	5.5	0.22	1.1
37.4	13.9	0.37	2.9
41.2	100.2	2.43	21
45	167.9	3.73	35.1
151.4	181	1.20	37.9
210	100	0	0
210	100	0	0
total ug:	14.35		

LabChip analysis of 20 percent of the purification allowed for visualization and quantitation of high stoichiometry components from the mixture. Electropherogram, pseudo-gel, and quantitation results are shown.

rpoA-6xHIS/V5 isolate 2

Size [kDa]	Rel. Conc. [ug/ml]	Rel. molarity [uM]	% Total
3.2	0	0	0
4.1	0	0	0
6	0	0	0
9.3	0	0	0
14.1	14.6	1.04	3.4
16.4	49	2.99	11.3
19.9	17.3	0.87	4
24	10.5	0.44	2.4
38.3	23.7	0.82	5.4
41.9	118.1	2.82	27.2
45.5	181	3.54	37.7
152.5	40.9	0.27	9.4
210	100	0	0
total ug:	13.07		

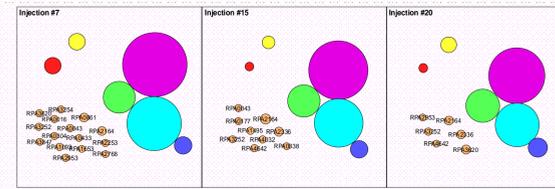
Results: Carryover and Systematic Performance Degradation

One big issue with attempting to perform large-scale analyses of protein complexes is the problem of sample carryover. Although we ran a "blank" sample between each, we observed a considerable amount of RPA2164 in our RPA3226 analyses. Because it varied relative to the other constituents in the affinity purification, some or all of this signal was due to sample carryover.

For this particular set of experiments, we noted a slight but systematic degradation in the performance of our analysis. We are currently working to identify the relative contributions of sample loss in the autosampler vials, column degradation, and other instrument related issues.

1/10 Samples

~ 5 ug total protein

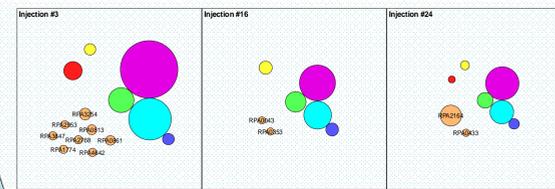


Each panel is a graphical representation of the results, with the area of each circle representing the total number of spectra that matched a particular protein. All proteins with 2 spectral hits or greater have been included. Color codes are as per the RNA polymerase diagram at the bottom middle of the figure (Beta - purple, Beta' - light blue, Alpha - green, et cetera). Actual injection numbers are indicated in the upper left-hand corner of each panel.

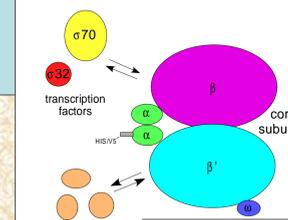
A composite "average" of the three replicates represent all proteins with at least one peptide found in 2 of 3 replicates. Here, the area of each circle represents the average number of spectra collected for a protein. Non-core protein constituents are only labeled if they were found in both pull-down experiments.

1/100 Samples

~ 500 ng total protein



R. palustris RNA polymerase



Proteins expected to be observed in affinity purifications of RNA polymerase. Protein colors and relative positions are maintained in the graphs of the sample replicates and experimental "averages".

RPAT288	10	6.9	5.3	16.9	9.9
K.LIGIIEVLEK +1	3.3				
K.LIGIIEVLEK +2	3.9				
K.LIGIIEVLEK +3	3.9				
K.LIGIIEVLEK +4	4.0	5.8	6.1	4.7	3.8
K.LIGIIEVLEK +5	3.3				
K.LIGIIEVLEK +6	3.3				
K.LIGIIEVLEK +7	3.5				
K.LIGIIEVLEK +8	3.8				
K.LIGIIEVLEK +9	3.0	2.9	3.2	6.2	6.2
K.LIGIIEVLEK +10	6.6	6.1	6.2	6.1	
K.LIGIIEVLEK +11	6.5	6.0	6.1	6.0	6.0
K.LIGIIEVLEK +12	3.1				
K.LIGIIEVLEK +13	3.0	3.4	3.1	3.3	

RPAT326	33.9	46	40.1	49	37.2	44.3
K.LIGIIEVLEK +1	4.5					
K.LIGIIEVLEK +2	4.5					
K.LIGIIEVLEK +3	3.7	3.3	4.0	4.0	3.1	3.0
K.LIGIIEVLEK +4	3.3					
K.LIGIIEVLEK +5	3.4	3.1	3.2	3.5	3.2	3.5
K.LIGIIEVLEK +6	3.1					
K.LIGIIEVLEK +7	3.1					
K.LIGIIEVLEK +8	3.1					
K.LIGIIEVLEK +9	3.1					
K.LIGIIEVLEK +10	3.1					
K.LIGIIEVLEK +11	3.1					
K.LIGIIEVLEK +12	3.1					
K.LIGIIEVLEK +13	3.1					

A comparison of the different peptides identified across each of the replicates for two of the proteins, sigma70 (yellow) and rpoA (green). Numbers in the first row indicate percent sequence coverage for that protein. Numbers in peptide rows report SEQUEST Xcorr for that particular peptide.

Results: Peptide Level Consistency

While the results from sample to sample were consistent with number of peptide identifications (also percent sequence coverage), the specific peptides which were identified varied far more. For the tagged protein itself (RPA3226), in the 1/10 dilution only 12 out of the 33 total identified peptides were found in every sample and in the 1/100 dilution only 3 out of 15. If the analysis depends on identifying precisely the same peptides each time, some combination of increased protein amounts and increased depth of analysis (a la MudPIT) will be required.

CONCLUSION

- In this case the apparent biological variation in the purification did not seem to reflect on the rest of the analysis (i.e. small variations in individual protein amounts did not seem to affect qualitative protein identifications).
- If the goal of the analysis is limited to identification of high-stoichiometry proteins within a complex, it appears that approximately 500 ng of total protein is sufficient. However, if the goal is to identify sub-stoichiometric or lower affinity members it will be necessary to operate at the ~5 ug range or higher.
- While the total peptides or spectra identified for a particular protein were consistent from sample to sample, the specific identified peptides were not. If such consistency were needed, a more rigorous analytical approach would be needed.
- These data should be useful as we tune our pipeline for high throughput versus depth of analysis. These data will also serve as a baseline for instrumentation and workflow development and for the meta-analysis that attempts to define bona fide biological complexes from these data.

REFERENCES

- Gavin, A.C., et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes*. Nature, 2002. 415(6868): p. 141-7.
- Ho, Y., et al., *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry*. Nature, 2002. 415(6868): p. 180-3.
- Butland, G., et al., *Interaction network containing conserved and essential protein complexes in Escherichia coli*. Nature, 2005. 433(7025): p. 531-7.
- Rigaut, G., et al., *A generic protein purification method for protein complex characterization and proteome exploration*. Nat Biotechnol, 1999. 17(10): p. 1030-2.
- MacCoss, M.J., et al., *Shotgun identification of protein modifications from protein complexes and lens tissue*. Proc Natl Acad Sci U S A, 2002. 99(12): p. 7900-5.
- Tabb, D.L., W.H. McDonald, and J.R. Yates, 3rd, *DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics*. J Proteome Res, 2002. 1(1): p. 21-6.
- Tabb, D.L., et al., *DBDigger: Reorganized Proteomic Database Identification That Improves Flexibility and Speed*. Anal Chem, 2005. 77(8): p. 2464-2474.

ACKNOWLEDGMENTS

This research was sponsored by the Office of Biological and Environmental Research, U.S. Department of Energy. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.