# Technology Developments for Characterizing the Complex Protein Datasets of Microbial Communities

Christine Shook[1,3]; Melissa Thompson[2,3]; Nathan VerBerkmoes[3]; Abijeet Borole[3]; Manesh Shah[3]; Frank Larimer[3]; Brian Davison[3]; Robert Hettich[3]

[1]University of Tennessee, Knoxville, TN; [2]ORNL-UTK GST Graduate School, Oak Ridge, TN; [3]Oak Ridge National Lab, Oak Ridge, TN

## OVERVIEW

- **Determining the physiological states of microbes within complex communities and environments is an important analytical and biological problem for MS-based proteomics.**
- **Recent studies have shown that the genomic sequence of major microbial constituents of communities can be mapped out in both simple microbial communities (Tyson, Nature 2004) and more complex environmental samples (Venter, Science 2004).**
- **The next step is characterizing protein complexes and mapping the proteome of both the individual organism and the organism as a member of a community combining community genomics and proteomics for a proteogenomic approach (Ram, Science 2005).**
- **"Shotgun" proteomics, directly assessing the protein composition of complex mixtures, is currently the method of choice.**
- **The goal of this study is to assess and improve MS-based "shotgun" proteomics technology for analyzing simple known microbial mixtures. The purpose is to improve current techniques and develop analytical procedures for analyzing highly complex communities and environmental samples.**

## INTRODUCTION

Recently, there have been strong efforts to develop techniques for genomic sequencing and annotation of microbial communities (metagenomics). With the potential of partial or near complete microbial genomes obtained from environmental samples, along with the rapid proliferation of isolated microbial genomes, systems biology in microbial communities by combining genomic, transcriptomic, proteomic and metabolic studies may be possible in the near future. Our current studies seek to develop and demonstrate metaproteome analysis techniques for mixed cultures, establishing the basis for "***whole community proteomics***". Currently our major focus is to use controlled, simple microbial mixtures of four species (*E. coli*, *S. cerevisiae*, *R. palustris*, and *S. oneidensis*) to develop MS-based proteomics methods, as well as the proteome bioinformatics tools for detailed analysis of sequenced microbial communities. Our goal is to provide "deep" and "wide" proteome measurements of complex microbial mixtures. Some of the considerations for any microbial community proteome project are highlighted below.

### Considerations for Proteome Analysis of Microbial Communities

- **Level of DNA sequence availability and quality of annotation for species in that community**
- **Diversity and dynamic range of species in the community**
- **Quantity and quality of community available (i.e. how much total protein can be obtained)**
- **Inter- and intra- species relationship at the amino acid level**

## EXPERIMENTAL

### Cell Growth and Production of Protein Fractions:

- The four microbes (**Figure 1**) used in this study (*E. coli*, *R. palustris*, *S. cerevisiae*, and *S. oneidensis*) were all grown individually and mixed after cell harvesting at appropriate concentrations based on wet cell weight. **Figure 2** illustrates the different mixtures analyzed in this study.
- *R. palustris* was used as the target organism and grown under three different metabolic states: photoheterotrophic, chemoheterotrophic, and nitrogen fixation and then mixed with the other three species at 25% (Mix 1). Comparisons were then made between **nitrogen fixation vs. photoheterotrophic** and **chemoheterotrophic vs. photoheterotrophic**.
- Mixed cell pellets were washed twice with Tris buffer and disrupted with sonication. A crude and membrane fraction were separated by ultracentrifugation at 100,000g for 1 hour. Protein fractions were quantified and digested with sequencing grade trypsin using standard protocol. The digested mixed proteomes were desalted via solid phase extraction and concentrated to ~10µg/µL and frozen at -80ºC until analysis.

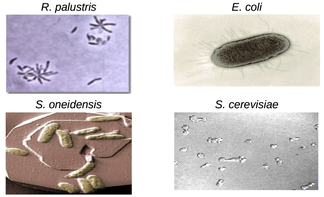



**Figure 1** Test species for Artificial Mixture

| Organism | Mix 1 | Mix 2 | Mix 3 | Mix 4 | Mix 5 |
|---|---|---|---|---|---|
| *E. coli* | 25% | 32% | 33% | 33% | 33% |
| *S. cerevisiae* | 25% | 32% | 33% | 33% | 33% |
| *S. oneidensis* | 25% | 32% | 33% | 33% | 33% |
| *R. palustris* | 25% | 5% | 1% | 0.2% | 0.0% |

**Figure 2** Concentrations in Artificial Mixtures

### LC-MS/MS Analysis and Database Searching:

- The tryptic digestions of the crude and membrane fractions from all mixtures were analyzed via 24-hour two-dimensional LC-MS/MS with a split-phase MudPIT column as described in McDonald, W.H., IJMS, 2002. The LC-MS system was composed of an Ultimate HPLC (LC Packings) and an LCQ-DECA XP ion trap mass spectrometer (Thermo Finnigan) or a Surveyor pump (Thermo Finnigan) and LTQ linear trapping quadrupole (Thermo Finnigan). Each analysis required ~500µg of starting protein material from either the crude or membrane fraction.
- The resultant MS/MS spectra from the mixed proteomes were searched with SEQUEST (Thermo Finnigan) against the four databases highlighted below. The resultant files were filtered and sorted with DTASelect (Tabb, Journal of Proteome Research 2004) and unique and non-unique peptide identifications were extracted with in-house perl scripts. Reverse database searches were employed to test false positive levels.

### Databases Used to Search MS/MS Spectra

- **DB4** - the four species in the mixtures:
  - *E. coli*
  - *R. palustris*
  - *S. cerevisiae*
  - *S. oneidensis*
- **DB13** - the four species plus 1 plant and 8 other microbes
- **DB large** - 261 species with 1,011,612 total proteins
- **DB4 rev** - DB4 plus DB4 in reversed sequence

## RESULTS AND DISCUSSION

### Initial Mixture Studies:

- The initial experiments for this study have focused on testing current 2D-LC-MS/MS methodologies to analyze an artificial 4 microbe mixture of *E. coli*, *R. palustris*, *S. cerevisiae* and *S. oneidensis*.
- Two aspects of community proteomics were analyzed in this study:
  - To determine at what level functionally meaningful results could be obtained from a target species (*R. palustris*) whose concentration was decreased over a range of 25% to 0.2%. All experiments were conducted with current 2D-LC-MS/MS technologies.
  - To test the impact of database size on the identification of *unique peptides* from the four microbial species in the sample. *Unique peptides are defined as those peptides unique to a given protein and specific to a species in a given database.*
- **Figure 3** highlights the results from the 4 species database with decreasing concentrations of *R. palustris*.
- **Figure 4** highlights the results from the 13 species database with decreasing concentrations of *R. palustris*.
- **Figure 5** highlights results from the 261 species database of the 25% mixture.

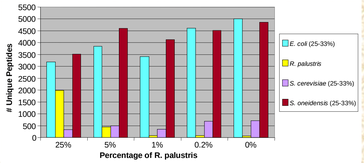



**Figure 3** Unique Peptides Identified with DB4

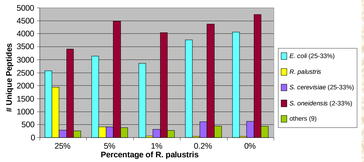



**Figure 4** Unique Peptides Identified with DB13

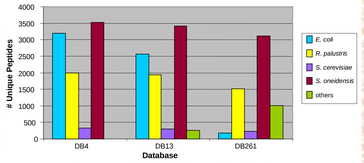



**Figure 5** Database comparison of the 25% mixture

### *R. Palustris* Proteome from Initial Mixture Studies

- The functional categories of the identified proteome from *R. palustris* analyzed from the different mixtures were compared with our *R. palustris* baseline proteome (VerBerkmoes et al. submitted J. Bacteriology). **Figure 6** illustrates this comparison: top-left pie chart is *R. palustris* baseline proteome and the other three pie charts are the 25%, 5%, & 1% mixtures.
- **Figure 7** highlights the results from a single protein identified in the mixtures. This protein puhA is a critical subunit of the photoreaction center and is indicative of photosynthesis.
- **Figures 8 and 9** illustrate diagnostic tandem mass spectra of peptides from puhA at the different concentrations of *R. palustris* in the sample, ranging down to 0.2%.

**Figure 6** ***R. palustris*** **Functional Categories**

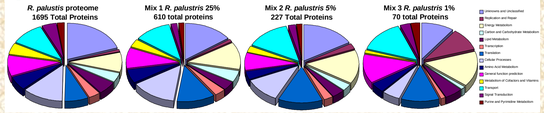



**Figure 7** PuhA, a Component of the Photosynthetic Reaction Center from ***R. palustris***

| | | | | |
|---|---|---|---|---|
| **Mix 1** | **44.70%** | | | |
| **rpal:1548** | **Charge State** | **Xcorr** | **DelCN** | **Peptide** |
| * | 3 | 4.1787 | 0.5624 | K.TVPSTSNDRPNVALTPAAPWPGAPFVPTGNPFADGVGPGSYAQR.A |
| * | 3 | 5.8592 | 0.4355 | R.ADVPELGLDNLPIIVPLR.A |
| * | 2 | 5.6884 | 0.5109 | R.ADVPELGLDNLPIIVPLR.A |
| * | 3 | 3.788 | 0.4484 | R.ADVPELGLDNLPIIVPLRAAK.G |
| * | 1 | 2.7334 | 0.2724 | R.YLEVEVAK.S |
| * | 2 | 4.0343 | 0.5105 | R.VILLPVPFALINDPFGK.V |
| * | 3 | 5.8444 | 0.5249 | R.VILLPVPFALINDPFGK.V |
| * | 2 | 3.2296 | 0.6166 | R.VILLPVPFALINDPFGKVSVDAIR.G |
| * | 2 | 4.5225 | 0.5718 | K.VSVDAIRGDQFAGVPTTSKGDQVSK.L |
| * | 3 | 3.8473 | 0.3707 | K.VSVDAIRGDQFAGVPTTSKGDQVSK.L |
| **Mix 2** | **42.70%** | | | |
| **rpal:1548** | **Charge State** | **Xcorr** | **DelCN** | **Peptide** |
| * | 2 | 2.5409 | 0.3636 | K.IGVPAPPDPK.T |
| * | 3 | 3.5846 | 0.4472 | K.TVPSTSNDRPNVALTPAAPWPGAPFVPTGNPFADGVGPGSYAQR.A |
| * | 3 | 5.8143 | 0.4663 | R.ADVPELGLDNLPIIVPLR.A |
| * | 2 | 5.6008 | 0.4809 | R.ADVPELGLDNLPIIVPLR.A |
| * | 2 | 4.1629 | 0.369 | R.ADVPELGLDNLPIIVPLRAAK.G |
| * | 2 | 2.9884 | 0.485 | R.VILLPVPFALINDPFGK.V |
| * | 2 | 3.7545 | 0.4465 | R.GDQFAGVPTTSKGDQVSK.L |
| **Mix 3** | **17.30%** | | | |
| **rpal:1548** | **Charge State** | **Xcorr** | **DelCN** | **Peptide** |
| * | 3 | 5.1517 | 0.583 | K.TVPSTSNDRPNVALTPAAPWPGAPFVPTGNPFADGVGPGSYAQR.A |
| **Mix 4** | **7.10%** | | | |
| **rpal:1548** | **Charge State** | **Xcorr** | **DelCN** | **Peptide** |
| * | 2 | 4.4881 | 0.4467 | R.ADVPELGLDNLPIIVPLR.A |

**Figure 8** Diagnostic Peptide at 1% ***R. palustris***

**MS/MS at 25% and 1% *R. palustris***

TVPSTSNDRPNVALTPAAPWPGAPFVPTGNPFADGVGPGSYAQR

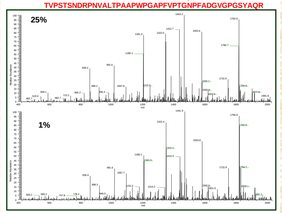
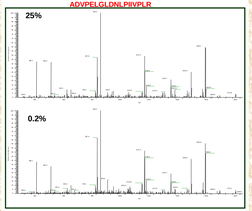



**Figure 9** Diagnostic Peptide at 0.2% ***R. palustris***

**MS/MS at 25% and 0.2% *R. palustris***

ADVPELGLDNLPIIVPLR




### Comparison of *R. palustris* Metabolic States in a Simple Microbial Mixture

- Experiments were performed to test the ability to detect differential metabolic states of R. palustris in the presence of the three other microbes (all at 25%):
  - Aerobic (Chemoheterotrophic)
  - Anaerobic (Photoheterotrophic)
  - Nitrogen Fixation
- **Table 1a**: number of proteins identified at the 1- and 2-peptide level in all three growth states
- **Table 1b**: proteins identified for each species in the LTQ Nitrogen Fixation experiment
- **Table 2** number of unique peptides found in the nitrogen fixation state with the associated percentage of false positives identified using the reverse database (DB4 rev). Results for the other states were similar.
- **Table 3**: Differential expression of some nitrogen fixation proteins from the nitrogen fixation vs. photoheterotrophic metabolic states.
  - strongest differences highlighted in red
  - blue highlighting represents known nitrogen fixation proteins
- **Table 4** highlights unique peptides detected from the NifK beta-chain protein.
- *Requirements for differentially expressed proteins*: at least 4 unique peptides or more than 30% sequence coverage difference

**Table 1a** Protein Totals with Conservative and Liberal Filters

| Metabolic State | Experiment | 1 Peptide | 2 Peptide |
|---|---|---|---|
| | LCQ Run 1 | 2679 | 1448 |
| Photoheterotrophic | LCQ Run 2 | 2593 | 1357 |
| | LTQ | 5391 | 3339 |
| | | | |
| | LCQ Run 1 | 2615 | 1335 |
| Chemoheterotrophic | LCQ Run 2 | 2306 | 1180 |
| | LTQ | 5517 | 3354 |
| | | | |
| | LCQ Run 1 | 2588 | 1420 |
| Nitrogen Fixation | LCQ Run 2 | 2164 | 1177 |
| | LTQ | 5762 | 3650 |

**Table 1b** Species Protein Counts for Nitrogen Fixation LTQ

| | | |
|---|---|---|
| Nitrogen Fixation | *E. coli* | 1182 |
| | *R. pal* | 1274 |
| *LTQ Breakdown (2 peptide)* | *S. oneidensis* | 763 |
| | *S. cerevisiae* | 431 |
| | Total | 3650 |

**Table 2** Unique Proteins and False Positives Identified

| N2 Fixation Peptides | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Run 1 | | | Run 2 | | | LTQ | | |
| 1 pep | 2 pep | | 1 pep | 2 pep | | 1 pep | 2 pep | |
| 9235 | 8173 | Unique | 7754 | 6844 | Unique | 27775 | 25617 | |
| 310 | 13 | Rev Unique | 274 | 16 | Rev Unique | 1126 | 157 | |
| 3.2 | 0.15 | % false pos. | 3.4 | 0.23 | % false pos. | 3.9 | 0.6 | |

| N2 Fixation Proteins | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Run 1 | | | Run 2 | | | LTQ | | |
| 1 pep | 2 pep | | 1 pep | 2 pep | | 1 pep | 2 pep | |
| 2461 | 1399 | Unique | 2072 | 1162 | Unique | 5715 | 3557 | |
| 305 | 8 | Rev Unique | 266 | 8 | Rev Unique | 1047 | 78 | |
| 11.0 | 0.56 | % false pos. | 11.3 | 0.6 | % false pos. | 15.4 | 2.1 | |

**Table 3** Nitrogen Fixation vs. Photoheterotrophic Activated Genes

| UP_Nitrogen Fixation | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Anaerobic #1 | Anaerobic #2 | Anaerobic LTQ | N2 #1 | N2 #2 | N2 LTQ | |
| | LCQ 1 | LCQ 2 | LTQ | LCQ 1 | LCQ 2 | LTQ | Function |
| Gene | % Coverage | % Coverage | % Coverage | % Coverage | % Coverage | % Coverage | |
| RPA0774 | 0 | 0 | 0 | 38.4 | 38.4 | 71.4 | glnK2 GlnK, nitrogen regulatory protein P-II |
| RPA1927 | 0 | 0 | 0 | 32 | 32 | 57.7 | hypothetical protein |
| RPA1928 | 0 | 0 | 0 | 26.5 | 26.5 | 61.5 | ferredoxin-like protein [2Fe-2S] |
| RPA2125 | 0 | 26.4 | 31.8 | 31.8 | 68.2 | 85.1 | conserved unknown protein |
| RPA2963 | 0 | 0 | 9.1 | 0 | 0 | 45.5 | leuA SOS response transcriptional repressor; leuA |
| RPA3011 | 0 | 0 | 5.9 | 0 | 0 | 58.5 | unknown protein |
| RPA3869 | 0 | 0 | 0 | 28.6 | 28.6 | 45.2 | amino acid uptake ABC transporter periplasmic solute-binding protein |
| RPA4286 | 0 | 0 | 0 | 12.5 | 12.5 | 21.5 | glnA II glutamine synthetase II |
| RPA4402 | 0 | 0 | 15.4 | 15.4 | 15.2 | 15.4 | fixA ferredoxin like protein; fixA |
| RPA4403 | 0 | 0 | 0 | 23.4 | 23.4 | 32 | fixC nitrogen fixation protein;fixC |
| RPA4404 | 0 | 0 | 0 | 13.6 | 13.6 | 25.5 | fixD electron transfer flavoprotein alpha chain protein fixD |
| RPA4405 | 0 | 0 | 0 | 13.9 | 13.9 | 59.8 | fixA electron transfer flavoprotein beta chain fixA |
| RPA4406 | 0 | 0 | 0 | 0 | 11.2 | 11.2 | nifS2 nitrogenase cofactor synthesis protein nifS |
| RPA4814 | 0 | 0 | 0 | 14.9 | 14.9 | 46.4 | ZnR ZnR |
| RPA4815 | 0 | 0 | 0 | 5.3 | 0 | 33.3 | nifH nitrogenase molybdenum iron protein nifK |
| RPA4818 | 0 | 0 | 0 | 45.9 | 45.9 | 65.1 | nifK nitrogenase molybdenum iron protein beta chain; nifK |
| RPA4819 | 0 | 0 | 0 | 37 | 37 | 63.2 | nifD nitrogenase molybdenum iron protein alpha chain; nifD |
| RPA4820 | 0 | 0 | 0 | 49.3 | 49.3 | 60.7 | nifH4 nitrogenase iron protein; nifH |
| RPA4823 | 0 | 0 | 0 | 68.2 | 68.2 | 69.3 | fixU; nifT conserved hypothetical protein |

**Table 4** Unique Peptides Identified in NifK

rpal:RPA4818

nifK nitrogenase molybdenum iron protein beta chain

| Unique Peptides Identified | Anaerobic_1 | Anaerobic_2 | Anaerobic_LTQ | N2_1 4h,5 | N2_2 4h,5 | N2_LTQ 6h,1 |
|---|---|---|---|---|---|---|
| K.AGDPLGIAYFAAVVSPEK.T +1 | | | | 4.0277 | 4.0277 | 3.757 |
| K.AGDPLGIAYFAAVVSPEK.T +2 | | | | 5.5955 | 5.5955 | 5.587 |
| K.AGDPLGIAYFAAVVSPEK.T +3 | | | | | | 0.8967 |
| K.ATSRGRTTSK.T +1 | | | | 2.6922 | 2.6922 | 2.6922 |
| K.ATSRGRTTSK.T +2 | | | | | | 2.6921 |
| K.ATSRGRTTSK.T +3 | | | | | | 2.584 |
| K.DKDGVRGFFVPTVYATRPVSGHFDTIFDENTMK.G | | | | | | 4.2375 |
| K.DKDGVRGFFVPTVYATRPVSGHFDTIFDENTMK.G +1 | | | | | | 3.9058 |
| K.DSGKDAVTKVAEK.T +1 | | | | 2.309 | 2.309 | 2.309 |
| K.DSGKDAVTKVAEK.T +2 | | | | | | 3.2695 |
| K.GSVSPEFDVLTVAATRPAVSGHFDTIFDENTMK.G +1 | | | | 3.8937 | 3.8937 | 4.2917 |
| K.GSVSPEFDVLTVAATRPAVSGHFDTIFDENTMK.G +2 | | | | 4.0392 | 4.0392 | 4.196 |
| K.IFDSKGKVTKVAEK.T +1 | | | | | | 2.6947 |
| K.IFDSKGKVTKVAEK.T +2 | | | | | | 3.6147 |
| K.LDAFEREHK.A +1 | | | | | | 4.6274 |
| K.LDKIFDSKMKATFAGEK.T +2 | | | | | | 6.2682 |
| K.LDKIFDSKMKATFAGEK.T +3 | | | | | | 4.2381 |
| K.REHVELFREPEITEEIERE.K +2 | | | | | | 2.6116 |
| K.REHVELFREPEITEEIERE.K +3 | | | | | | 2.5116 |
| K.WRKVYYDWAEYTRK.Y +2 | | | | 2.4026 | 2.4026 | 3.1167 |
| K.WRKVYYDWAEYTRK.Y +2 | | | | 0.6822 | 0.6822 | 4.5149 |
| K.WRKVYYDWAEYTRK.Y +3 | | | | | | 4.6465 |
| K.SGTVEVSLER.K +1 | | | | | | 2.0003 |
| K.SGTVEVSLER.K +2 | | | | | | 2.1001 |
| K.TQTVSPGR.- +2 | | | | 1.8051 | 1.8051 | 1.8051 |
| K.TQTVSPGR.- +3 | | | | | | 3.3468 |
| K.TLATAGHDGEVVAFNRHPVDIAIITDR.F +2 | | | | 6.6073 | 6.6073 | 4.9667 |
| K.TLATAGHDGEVVAFNRHPVDIAIITDR.F +3 | | | | 6.1363 | 6.1363 | 7.546 |
| K.TLPFVRSSSDCVLH +1 | | | | | | 3.6814 |
| K.VSALLEEEPSHSDDLK +1 | | | | 3.0283 | 3.0283 | 2.8839 |
| K.VSALLEEEPSHSDDLK +2 | | | | 4.8325 | 4.8325 | 5.8605 |
| R.EAALVRRFAE.A +1 | | | | 3.2189 | 3.2189 | 3.1923 |
| R.EAALVRRFAE.A +2 | | | | | | 2.9541 |
| R.EKDYSELVERKR.R +1 | | | | | | 3.122 |
| R.FLGAVSR.I +1 | | | | 2.3524 | 2.3524 | 2.4086 |
| R.FLGAVSR.I +2 | | | | | | 3.6108 |
| R.HVREHPLGGFDGVFRLV +2 | | | | | | 4.6491 |
| R.HVREHPLGGFDGVFTVGNIR.E +2 | | | | 4.5387 | 4.5387 | 5.3752 |
| R.HVREHPLGGFDGVFTVGNIR.E +3 | | | | | | 3.6154 |
| R.LGRVFDERK.H +1 | | | | 1.9794 | 1.9794 | |
| R.LGRVFDERK.H +2 | | | | 2.9275 | 2.9275 | 3.3336 |
| R.LKVAGEEEAAARNEIK.K +1 | | | | | | 2.8371 |
| R.LLKKAEEEAAARNEIK.K +3 | | | | 3.9776 | 3.9776 | 3.8069 |
| R.LSVTTLGFPSRSFYDR.T +2 | | | | 5.2028 | 5.2028 | 5.134 |
| R.LSVTTLGFPSRSFYDR.T +3 | | | | 5.8487 | 5.8487 | 6.0253 |
| R.SVRYSYGSSNKVLVE.I +2 | | | | | | 3.9294 |
| R.SVRYSYGSSNKVLVE.I +3 | | | | | | 3.6294 |
| R.SRVSGYGSSNKVLVE.I +2 | | | | 4.5365 | 4.5365 | 4.153 |

## CONCLUSIONS

- **Significant *R. palustris* proteome measurements were possible at 25% and 5%, with 1% borderline. A few unique proteins could be detected confidently at the 0.2% level.**
- **The effects of database size on unique peptide identification were examined. Larger databases diminish identification of unique peptides for any given species, but species such as *E. coli* with many close relatives in the database are affected more dramatically. Databases should be limited to the community genome plus a small number of distracters.**
- **The false positive rate found using the reverse database was relatively low for the identified peptides and proteins at the 2-peptide level. Proteins identified at the 1-peptide level had a comparatively higher false positive rate, indicating the conservative 2-peptide filter was more accurate.**
- **The major focus of future experiments will be on instrumental and bioinformatic developments to increase the confidence and extent of protein identification. LC-MS techniques will be tested for proteome characterization in larger communities, with the ultimate goal of applying these techniques to complex community and environmental samples.**

## ACKNOWLEDGMENTS

- Dr. John Yates (Scripps) and Dr. David Tabb (ORNL) for DTASelect/Contrast
- Exploratory funding from the ORNL Director's Research and Development Program
- ORNL-UTK Genome Science and Technology Graduate School (for M. Thompson)
- Oak Ridge National Laboratory is managed and operated by the University of Tennessee-Battelle, L.L.C., under contract DE-AC05-00OR22725 with the U.S. Department of Energy.