

Label-Free Differential Analysis: An Iterative Approach to Increased Coverage, Improved Statistics and Results

Michael Athanas¹; Michael J. MacCoss²; Amol Prakash³; Lukas Käll²; Daniela Tomazela²; Brendan Maclean²; Taha Rezai³; Bryan Krastins³; David Sarracino³; Scott Peterman⁴; Alejandra Garces⁵; Sarah Fortune⁵; Mary F. Lopez³

¹VAST Scientific, Cambridge, MA; ²University of Washington, Seattle, WA; ³Thermo Fisher Scientific, Cambridge, MA; ⁴Thermo Fisher Scientific, Somerset, NJ; ⁵Harvard University, Boston, MA



Overview

Purpose: To demonstrate a laboratory and informatics workflow from discovery to targeted measurement

Methods: Data were acquired using high-resolution LC-MS/MS using hybrid linear ion trap (Orbitrap™) and triple quadrupole (Vantage) mass spectrometers.

Results: Discovery workflow validated known secretion mechanism proteins in tuberculosis (TB); targeted SRM methods provided quantitative results for specific peptide transitions.

Introduction

Label-free differential analysis is typically conducted by analyzing a comprehensive set of samples on a high-resolution mass spectrometer. Considerable effort is spent on sample preparation and data acquisition, however, the statistical differential analysis that follows can often lead to inconclusive results due to ambiguous MS/MS identifications, low sequence coverage for proteins of interest, and single peptide hit protein identifications. Further, the statistical analysis yields many peptide-like analytes that are differentially detected, but have no MS/MS information to confirm identity. These caveats can compromise any biological inference that might be derived as a result of the statistical differential analysis. SRM targeted analysis can provide a vehicle for the high-throughput verification of putative protein and peptide candidates identified in high-resolution LC-MS/MS differential expression experiments. The ability to mine differential expression discovery MS data for optimal SRM method development would facilitate the verification process. In this report, we describe a novel iterative label-free analysis workflow comprising components for core-differential analysis, LC-MS/MS identification refinement, proteotypic peptide verification and analyte exploration. We applied this iterative workflow to the analysis of *Mycobacterium tuberculosis* proteins involved in the ESX1 secretion system (1). In order to comprehensively identify substrates of the ESX1 secretion system, we coupled high-resolution LC-MS/MS with novel label-free differential analysis software to analyze Mtb mutant strains lacking the ESX1 locus (2).

Methods

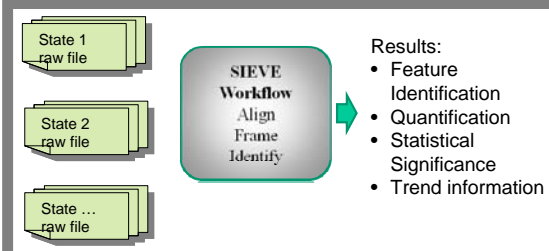
Samples

Mtb culture supernatants were prepared as described in (2). Secreted proteins were precipitated from the supernatants with TCA and the protein pellet was resuspended in SDS PAGE loading buffer. Samples were run approximately 1 cm into 10% SDS PAGE gels and the entire protein containing band was excised and subjected to in-gel digestion before loading onto the mass spectrometer.

Mass Spectrometry

High-resolution LC-MS/MS was run in a top 5 configuration at 60K resolution for a full scan, with monoisotopic precursor selection enabled, and with CID and HCD fragmentation modes on a Thermo Scientific LTQ Orbitrap hybrid mass spectrometer. LC-MS/MS data were analyzed with Thermo Scientific SIEVE software to determine differentially expressed peptides and proteins (Figures 1-7). SRM assays were developed on Thermo Scientific Vantage triple quadrupole mass spectrometer, Surveyor MS pump, Micro Autosampler, and IonMax source equipped with a low-flow metal needle. Thermo Scientific Pinpoint software was used to predict candidate peptides, choose multiple fragment ions for SRM assay design, build an instrument method and a sequence file, and to automatically confirm peptide identities and process quantitative data. Peptides were identified by co-eluting light and heavy transitions derived from synthetic peptide standards.

FIGURE 1. Data Processing – SIEVE is a label-free differential analysis tool for mass spectrometer data. Sets of biological or technical replicate data can correspond to “Control vs Treatment” experiment or a trend (time series, dosage study, biological category, etc.). The data are processed through the SIEVE workflow which consists of chromatographic alignment, frame discovery of potentially interesting features in the aggregate data set, and identification.



Analysis

Label-free differential analysis and protein identification were performed using the SIEVE™ v1.2 software. Within SIEVE, data are processed in three primary steps:

- **Alignment** – Full scan spectra from a designated reference measurement are compared with all other measurements. A correlation matrix is constructed from the comparison. An optimal path (dynamic programming) is extracted for correlation matrices constructed from comparing full scan spectra only.
- **Frame** – Potentially interesting features are exposed based upon high-intensity peaks found in the aligned collective data set. Individually, these peaks define frames *ie* well defined rectangular regions in the full scan (*m/z* versus retention time) plane
- **Identify** – After framing, MS2 fragment scans associated with each frame are processed with SEQUEST. Peptide quality scores are derived by SEQUEST processing against decoy shuffled database and processed using Percolator (3). Peptides assessed with less than 2% estimated false discovery rate are retained.

Post discovery, SIEVE results were imported into Pinpoint™ software. Five peptide transitions were selected for targeted measurement and acquisition methods were constructed for the Vantage™ triple quadrupole MS. Heavy labeled peptides (R or K) were synthesized and used to quantify peptide transition intensities.

FIGURE 2. Chromatographic alignment is based upon the pair wise MS full scan comparison of all experimental MS runs with respect to a chosen reference MS run.

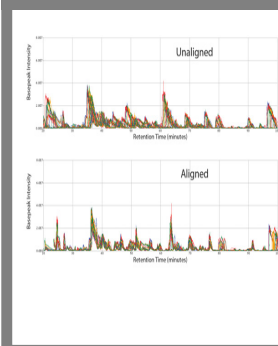


FIGURE 3. Overlapping correlation sub-matrices (tiles) are computed using a novel scalable adaptive tile algorithm. An optimal path through each tile is determined using dynamic programming and a final alignment score is calculated.

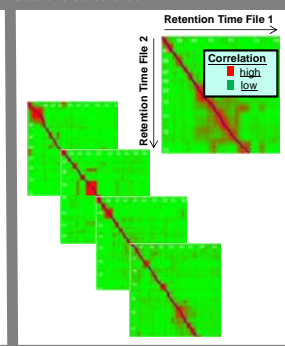


FIGURE 4. SIEVE Gel View - Frames (depicted as red rectangles in the plot below) have predefined dimensions in the *m/z* versus Retention Time plane. A frame represents a potentially interesting feature found in the collective data set.

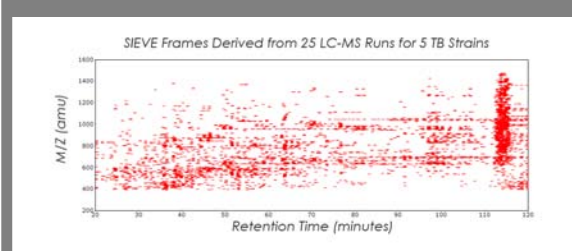


FIGURE 5. PCA of SIEVE Frames – Four technical replicate measurements were made on four mutant TB Strains (1, 20, 29, RD1) and the wild type control strain (RV). The ESX1 secretion mechanism is deleted in strain RD1 and disabled in 1. Strains 20 and 29 have other modifications.

PCA is an unsupervised clustering algorithm used to discover and reduce the dimensionality of a data set. The PCA calculation was performed on individual intensities for each frame. A natural clustering of the replicate measurements is depicted that reflects the nature of each TB strain.

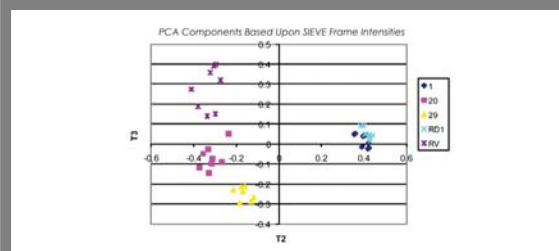


FIGURE 6. Protein Report – Peptide identifications are assigned using SEQUEST. Peptide quality scores are derived by SEQUEST processing against decoy shuffled databases and processed using Percolator (3). Peptides with an assessed 2% or less estimated false discovery rate are retained. Proteins->Peptides->MS2s are grouped in a hierarchy in the SIEVE protein report.

The SIEVE Frame shown below corresponds to the peptide LAGGVAVKA identified as Chaperonin GroEL [*Mycobacterium tuberculosis* H37] protein. This peptide is one of five selected for targeted selected reaction monitoring (SRM).

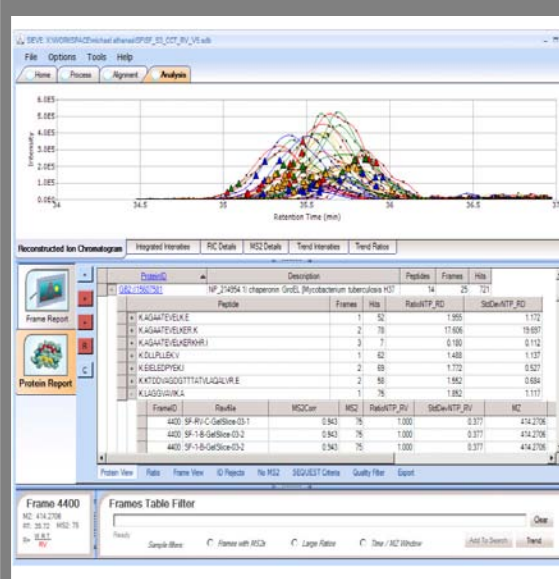


FIGURE 7. Ratios of wild type with respect to an RD1 restricted strain were calculated. Protein ratios were calculated using variance weighted averaging of each individual peptide measurement.

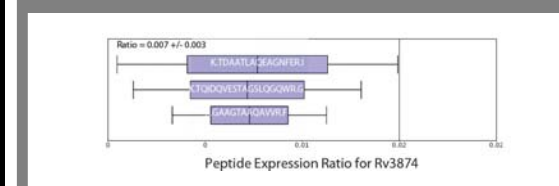


FIGURE 8. Suppression of several RD1 proteins responsible for attenuated virulence of BCG strain of mycobacteria were confirmed in the SIEVE analysis.

Locus	GID	Measured Ratio
Rv3875	57117165	0.003 +/- 0.001
Rv3874	15611010	0.007 +/- 0.003
Rv3616c	15610752	0.001 +/- 0.001
Rv3615c	15610751	0.001 +/- 0.001
Rv3865	15611001	0.409 +/- 0.060
Rv3870	15611006	0.001 +/- 0.001
Rv3871	15611007	0.001 +/- 0.001
Rv3877	15611013	0.295 +/- 0.064

FIGURE 9. Pinpoint Absolute Quantification / Calibration – Reverse calibration curves for individual transitions were constructed using the Pinpoint SRM analysis platform. Analysis is done on individual transition level. The level of detection (LOD) for the peptides was 500 attomoles and level of accurate quantitation (LOQ) was 1 femtomole on column. CV's for replicate samples were less than 20% for all target peptides.

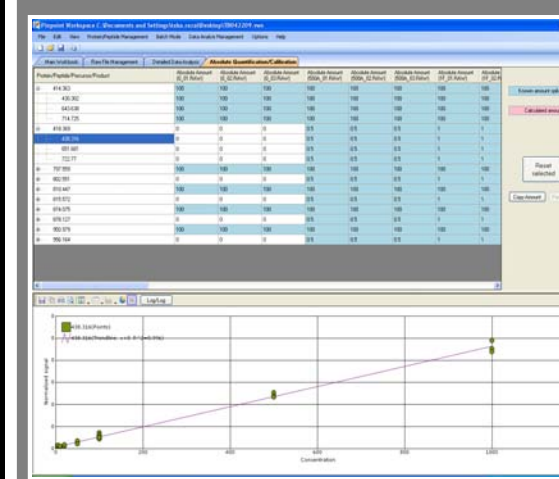
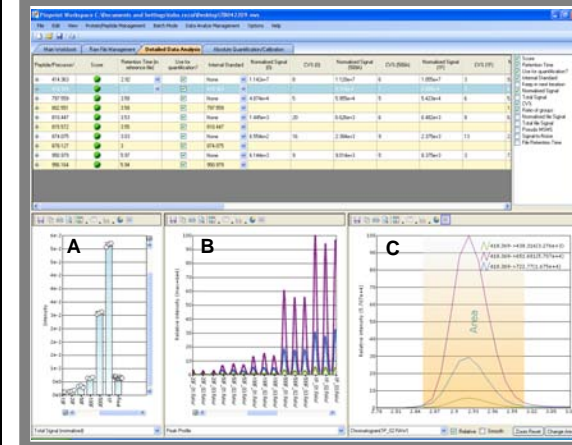


FIGURE 10. Pinpoint Detailed Data Analysis – The transitions corresponding to the five peptides of interest are assessed with Pinpoint. A) Numerical view of normalized signal intensities for each point on curve. B) Peak profile view of normalized signal intensities with overlaid SRM transitions shown in multicolor C) Isolated peak view of single sample illustrating multiple overlaid SRM transitions



Conclusions

- Replicate data derived from LC/MS acquisition were processed through the SIEVE workflow consisting of chromatographic alignment, feature determination (frames), and processing of frame associated MS2 fragments with SEQUEST.
- Subsequently, Percolator, a machine learning engine that trains on high-quality identifications, was able to drastically reduce false positive matches.
- The verified identifications were then imported into Pinpoint software to in order to develop optimized SRM assays for the target proteins.
- Using the approach described above, data were derived from substrates of *Mycobacterium tuberculosis* with and without the ESX1 secretion locus intact. Functional exploration of ESX1 is anticipated to provide insight into the multi-subunit cell envelop spanning structure.
- Results from the analysis confirmed the identification of the five previously identified secreted proteins as well as other differentially expressed proteins across the mutant strains.

References

1. Fortune SM, Jaeger A, Sarracino DA, Chase MR, Sasseti CM, Sherman DR, Bloom BR, Rubin EJ. Mutually dependent secretion of proteins required for mycobacterial virulence. *PNAS*, 2005 102:10676.
2. Abdallah M, Abdallah, et. al., "Type VII secretion – mycobacteria show the way", *Nature Reviews*, V5, November 2007, p883
3. Lukas Käll, Jesse Canterbury, Jason Weston, William Stafford Noble and Michael MacCoss. "Semi-supervised learning for peptide identification from shotgun proteomics datasets" *Nature Methods* 4:923 - 925, November 2007

SEQUEST is a registered trademark of the University of Washington. All trademarks are the property of Thermo Fisher Scientific Inc. and its subsidiaries. This information is not intended to encourage use of these products in any manners that might infringe the intellectual property rights of others.