

# Label-free differential analysis: An iterative approach to increased coverage, improved statistics and results

Michael Athanas<sup>1</sup>; Michael J. Maccoss<sup>2</sup>; Amol Prakash<sup>3</sup>; Lukas Kall<sup>2</sup>; Daniela Tomazela<sup>2</sup>; Brendan Maclean<sup>2</sup>; Taha Rezai<sup>3</sup>; Bryan Krastins<sup>3</sup>; David Sarracino<sup>3</sup>; Scott Peterman<sup>4</sup>; Alejandra Garces<sup>5</sup>; Sarah Fortune<sup>5</sup>; Mary F Lopez<sup>3</sup>

<sup>1</sup>VAST Scientific, Cambridge, MA; <sup>2</sup>University of Washington, Seattle, WA; <sup>3</sup>ThermoFisher Scientific, Cambridge, MA; <sup>4</sup>ThermoFisher Scientific, Somerset, NJ; <sup>5</sup>Harvard 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 Orbitrap and Vantage mass spectrometers.

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

## Introduction

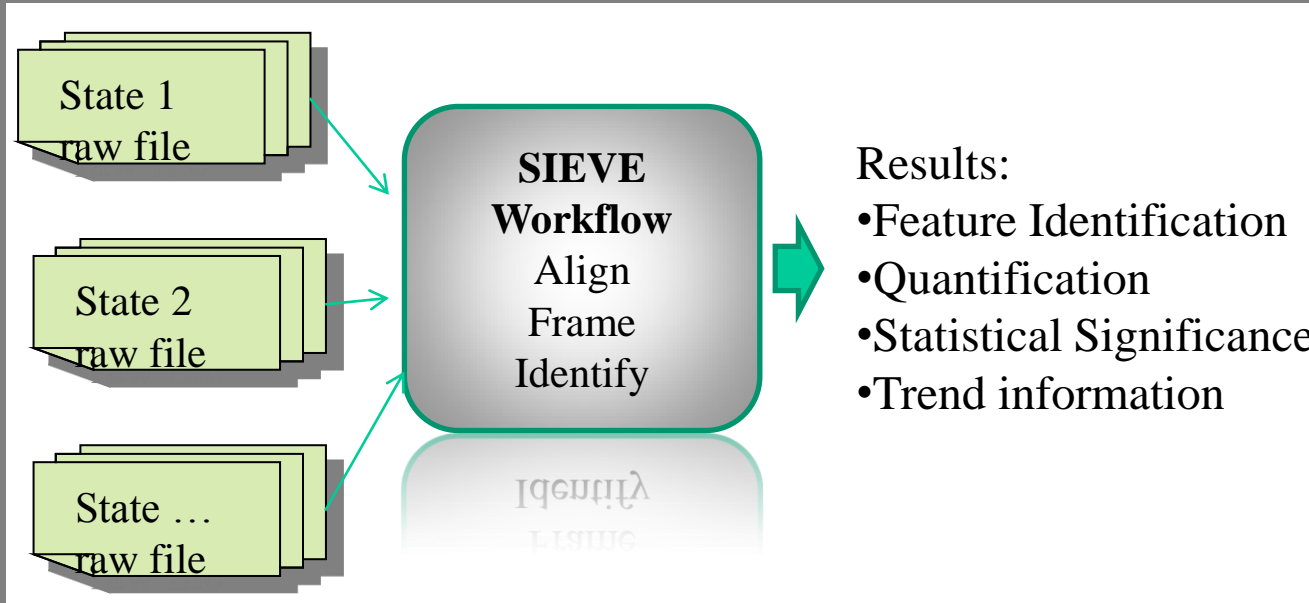
Label-free, differential analysis is typically conducted by analyzing a comprehensive set of samples on a high-resolution mass spectrometer. Considerable effort is spent on sample preparation and data acquisition, however, the statistical differential analysis that follows can often lead to inconclusive 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.

## Methods

Mtb culture supernatants were prepared as described in (1). 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 trypsin digestion before loading onto the mass spectrometer.

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 ThermoFisher Orbitrap. LC-MS/MS data were analyzed with SIEVE (ThermoFisher) to determine differentially expressed peptides and proteins (see Figure 1-7). SRM assays were developed on a ThermoFisher Vantage mass spectrometer, Surveyor MS pump, Micro Autosampler and an IonMax Source equipped with a low flow metal needle. Pinpoint software was used to predict candidate peptides and for choosing multiple fragment ions for SRM assay design, building an instrument method and a sequence file, and for automatic peptide identity confirmation and quantitative data processing. Peptides were identified by co-eluting light and heavy transitions derived from synthetic peptide standards.

**FIGURE 1. SIEVE Data Processing** – SIEVE is a label-free differential analysis tool for analyzing 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.



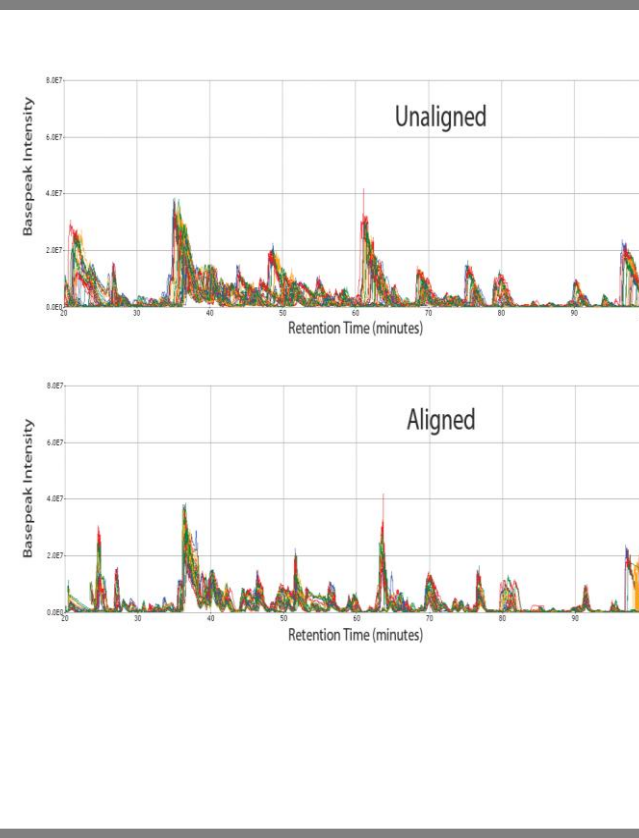
## Analysis

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

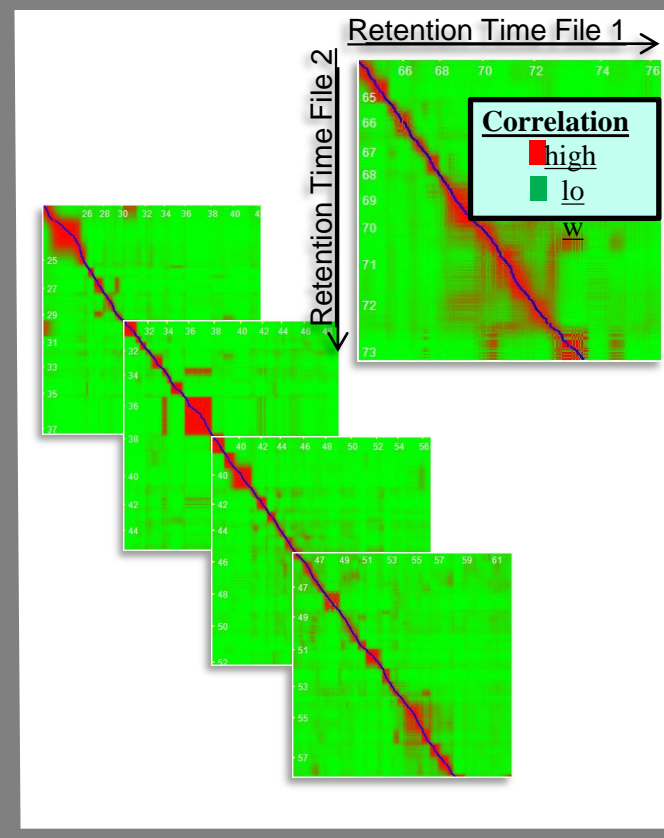
- Alignment** – Full scan spectra from a designated reference measurement are compared with all other measurements. A correlation matrix is constructed from the comparison. An optimal path (dynamic programming) is extracted for correlation matrices constructed from comparing full scan spectra only.
- 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
- Identification** – After framing, MS2 fragment scans associated with each frame are processed with SEQUEST. Peptide quality scores are derived by SEQUEST processing against decoy suffled database and processed using Percolator[2]. Peptides with an assessed 2% estimated false discovery rate are retained.

SIEVE results were imported into Pinpoint SRM software. Five peptide transitions were selected for targeted measurement. Acquisition methods were constructed for the TSQ triple quad. Synthetic heavy peptides were made and used to quantify peptide transition intensities.

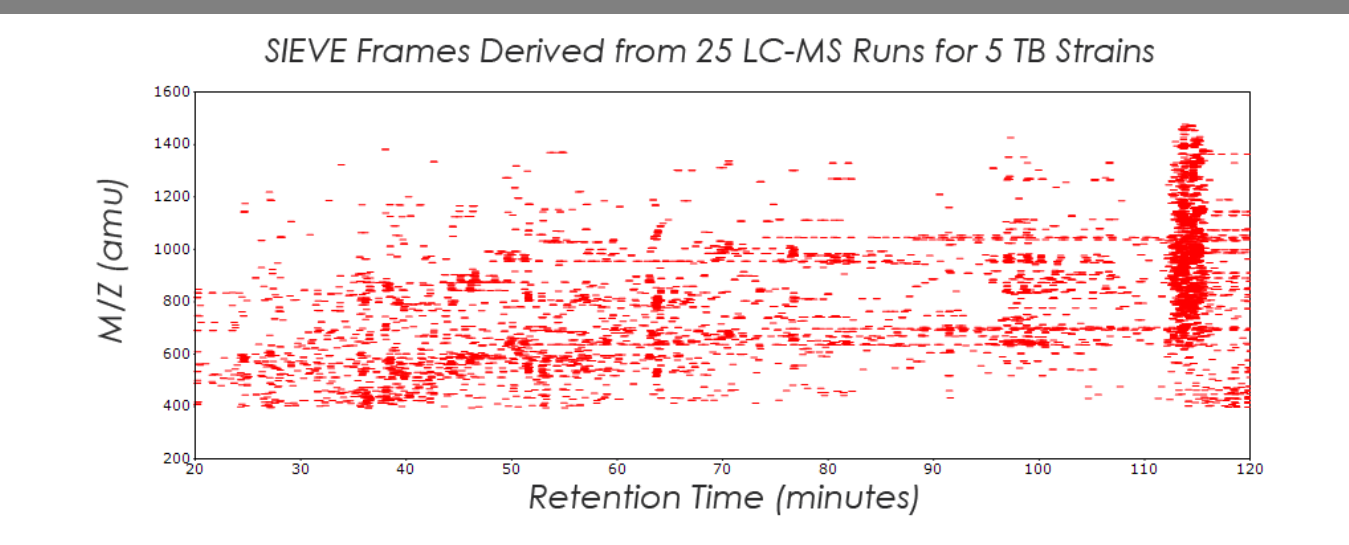
**FIGURE 2. Chromatographic alignment** is based upon the pairwise MS full scan comparison of all experimental MS runs with respect to a chosen reference MS run.



**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.

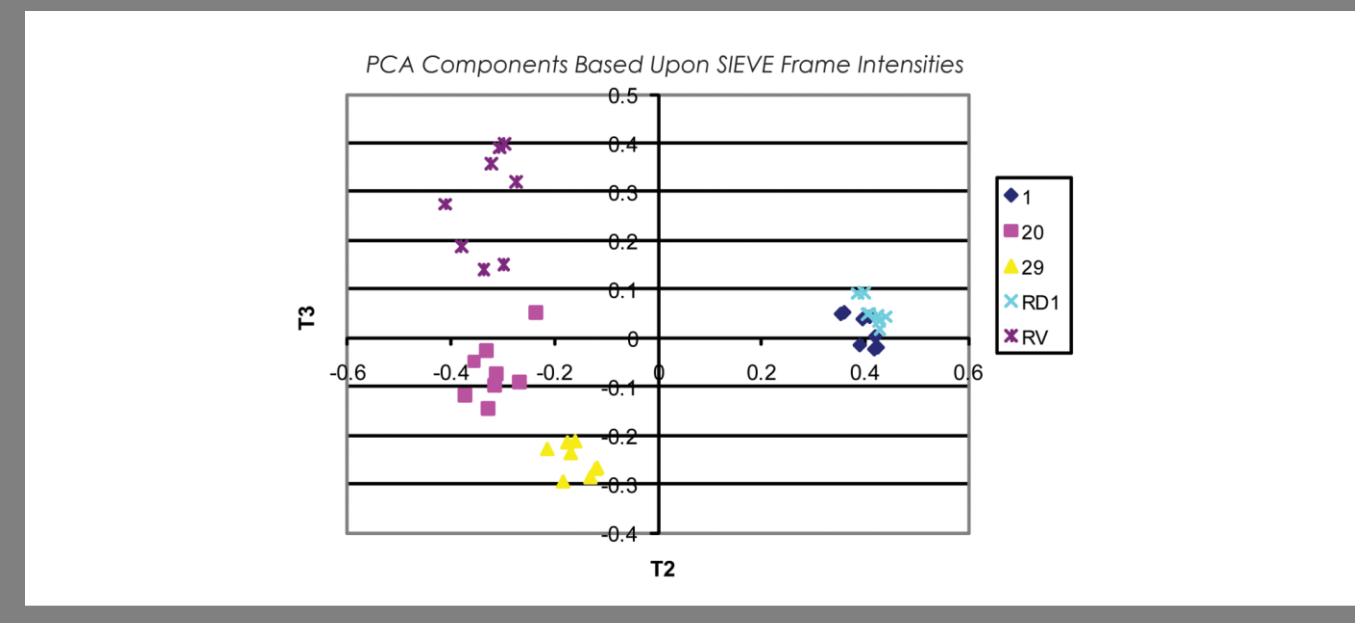


**FIGURE 4 SIEVE Gel View** - A frame (depicted as red rectangle in the plot below) had predefined dimensions in the M/Z versus Retention Time plane. A frame represents a potentially interesting feature found in the collective data set.



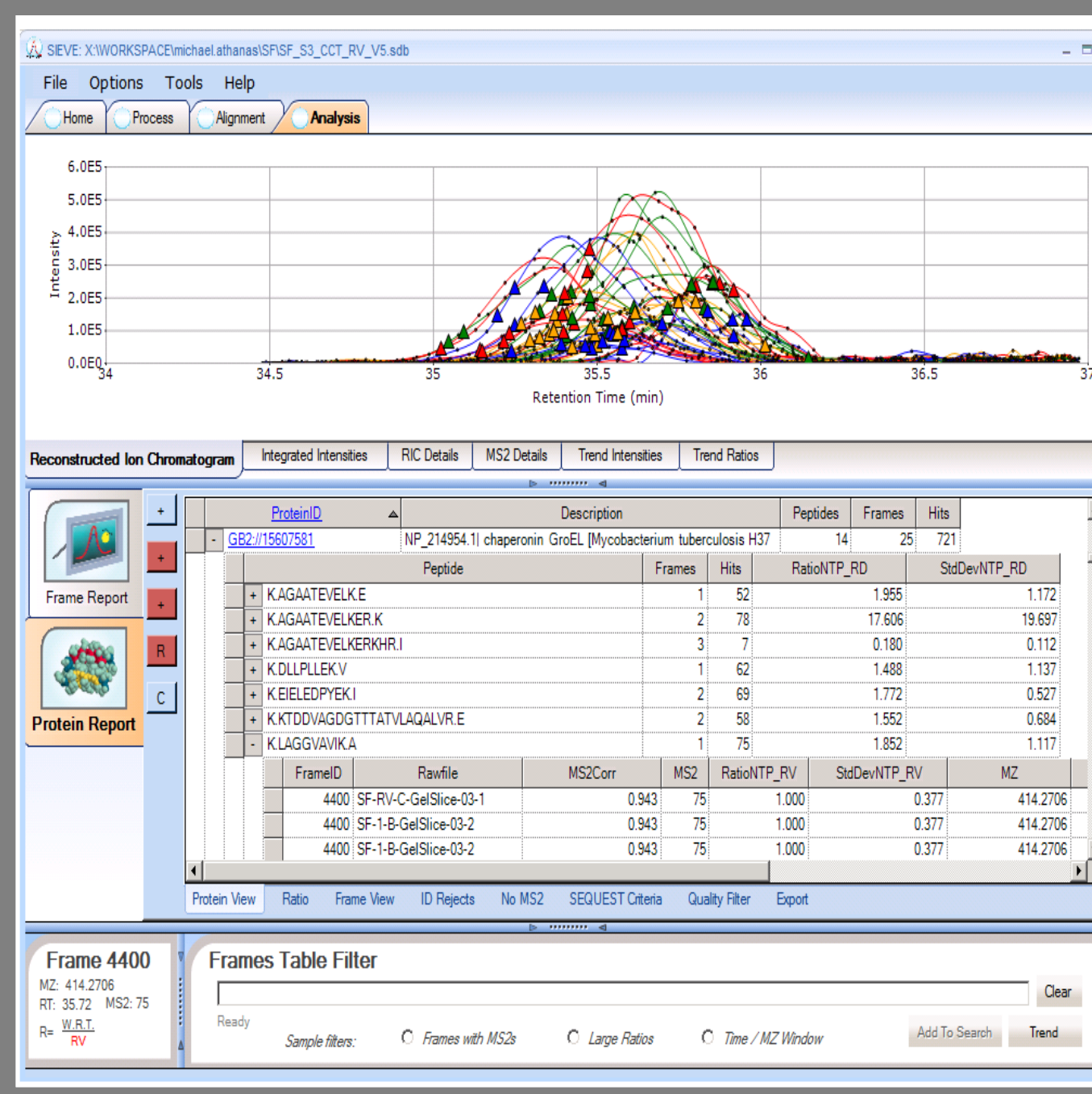
**FIGURE 5. PCA of SIEVE Frames** – Four technical replicate measurements were made on five mutant TB strains (1, 20, 29, RD1, RV). RV is the control strain. The ESX1 secretion mechanism is deleted in strain RV and disabled in 1. Strains 20 and 29 have other modifications.

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.

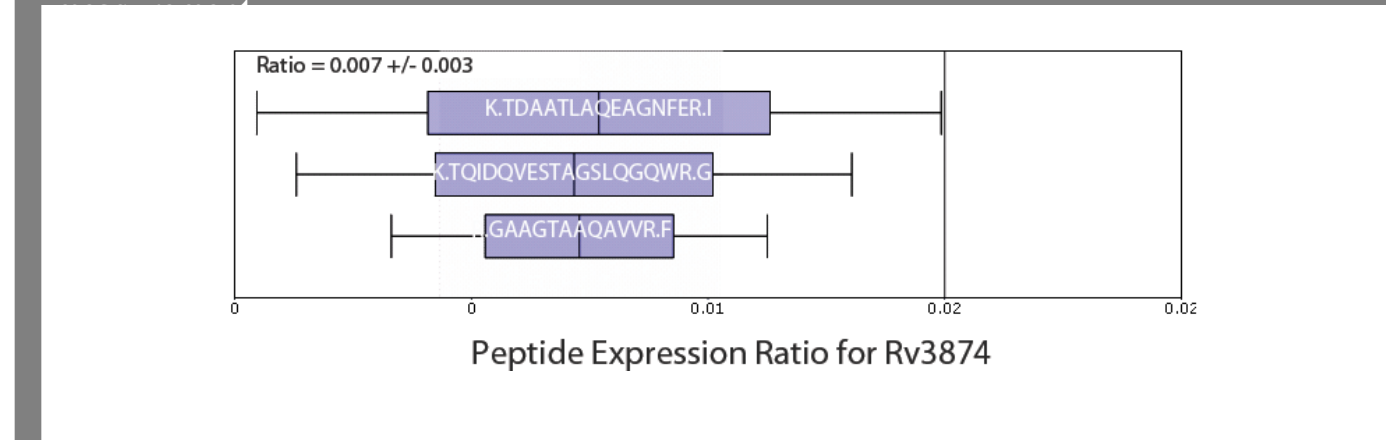


**FIGURE 6. Protein Report** – Peptide identification are assigned using SEQUEST. Peptide quality scores are derived by SEQUEST processing against decoy shuffled database and processed using Percolator[2]. Peptides with an assessed 2% estimated false discovery rate are retained. Proteins->Peptides->MS2s are grouped in a hierarchy in the SIEVE protein report.

The SIEVE Frame shown below corresponds to the peptide LAGGVAVIKA corresponding to the Chaperonin GroEL [Mycobacterium tuberculosis H37] protein. This peptide is one of five selected for targeted Selected Reaction Monitoring (SRM).



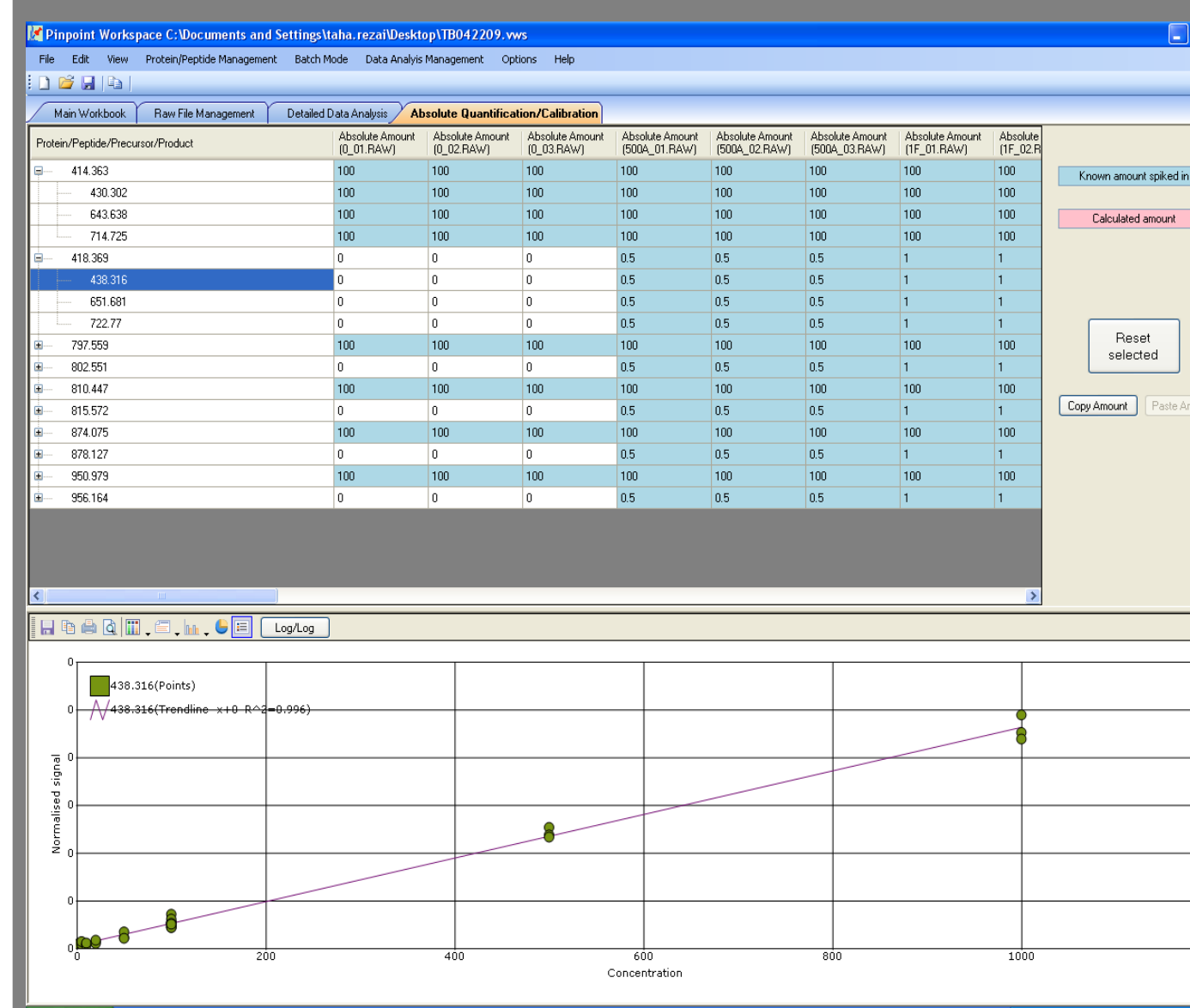
**FIGURE 7. Ratios** are calculated of wild type with respect to an RD1 restricted strain. Protein ratios are calculated using variance weighted averaging of each individual peptide



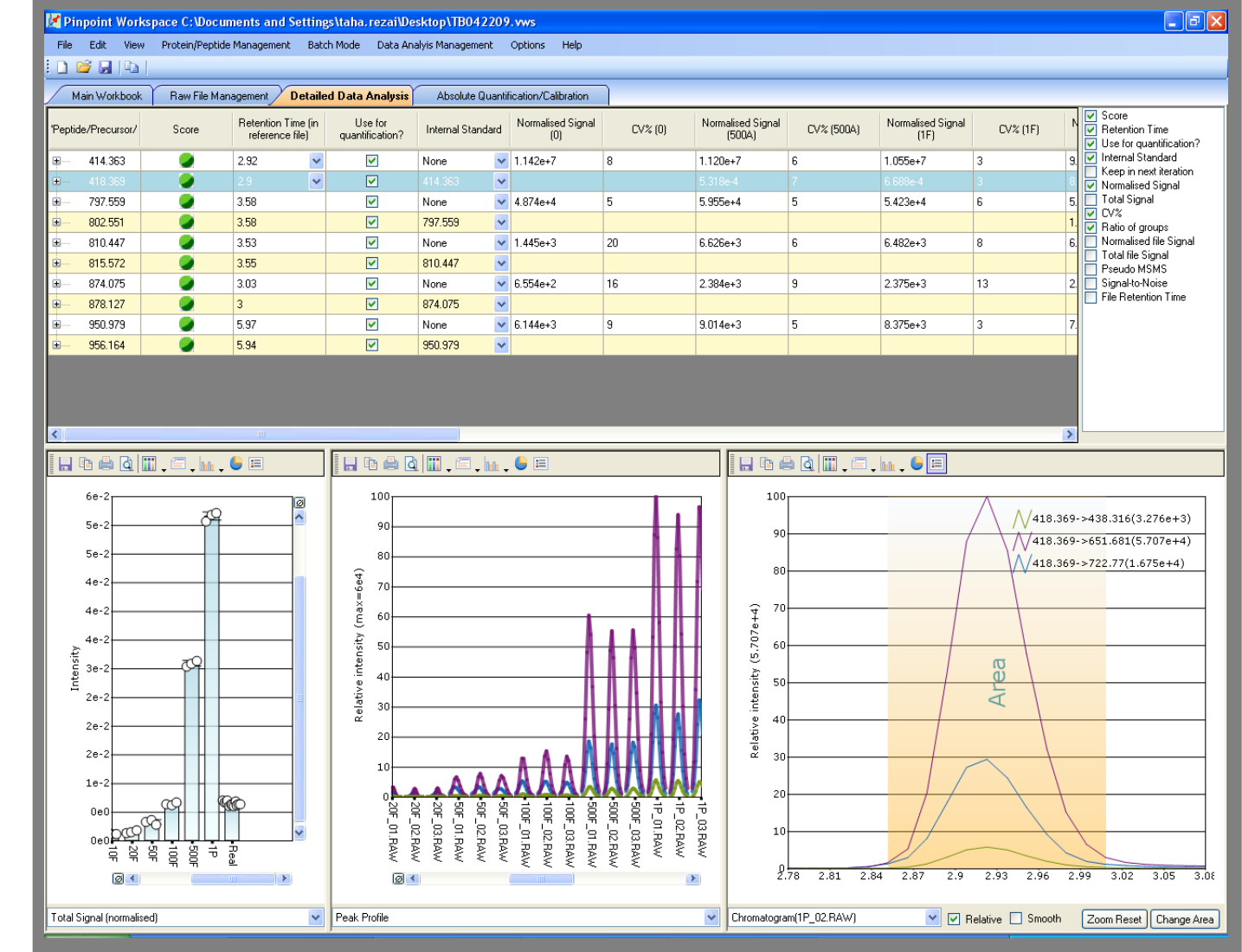
**FIGURE 8. Suppression** of several RD1 proteins responsible for attenuated virulence of BCG strain of mycobacteria are confirmed in the SIEVE analysis.

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. Known amounts for points on curve are used allowing unknown amounts in cell supernatant to be calculated against calibrations curve. Analysis is done on individual transition level.



**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

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. 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 are then fed into SRM Workflow software to list proteotypic peptides for all proteins with a goal to increase sequence coverage. 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

- Abdallah M. Abdallah, et al., “Type VII secretion – mycobacteria show the way”, Nature Reviews, V5, November 2007, p883
- 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

## Acknowledgements

We would like to thank Professor Sarah Fortune, Alejandra Garces, and Michael Chase from the Harvard School of Public Health for providing TB samples for this analysis.

SEQUEST is a registered trademarks of the University of Washington.