Supporting Information for: Probability Assignment and Protein Inference for Accurate Mass and Time Analysis

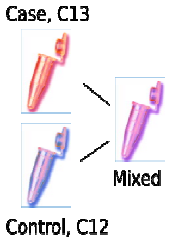
Damon May, Yan Liu, Wendy Law, Matt Fitzgibbon, Hong Wang, Sam Hanash, and Martin McIntosh

This document contains supporting information for our manuscript, particularly several figures that we were not able to put into the manuscript itself due to space considerations. These figures are generally associated a particular part of the manuscript, so we have organized this supporting information similarly to the manuscript. All plots are generated automatically by msInspect/AMT.

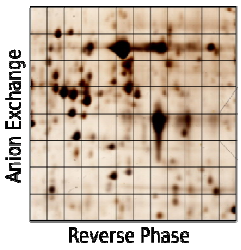
This document is delivered as part of a Supplementary Materials bundle that includes scripts that demonstrate the msInspect/AMT matching functionality. Please see the file README.txt for an overview of the bundle contents.

Introduction

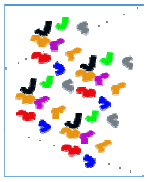
The following diagram describes the full experimental workflow, from sample preparation through protein inference, described in the manuscript. AMT data are combined with LC-MS/MS data to enhance the performance of an otherwise typical LC-MS/MS-based workflow.



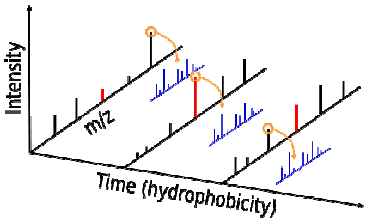
1. Deplete, Label, and Mix Case and Control



2. Separate Proteins

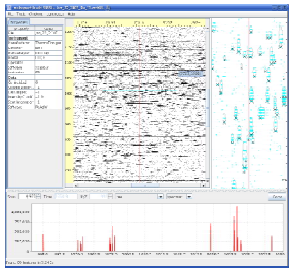


3. Digest Proteins



4. Acquire High-Resolution MS and MS/MS Data

Sample Processing

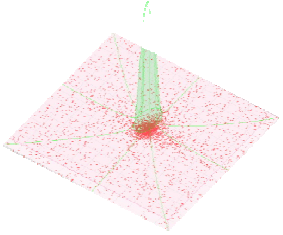


5. Locate Peptides in MS1, Identify Case/Control and Quantify

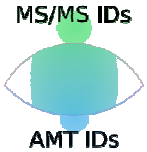
LC-MS1-Based Analysis

6. Search Protein DB to Derive Putative Peptide Sequences

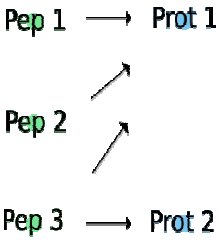
MS/MS-Based Analysis



7. Create AMT Database and Match MS1 Peptides, Assign IDs and Probabilities



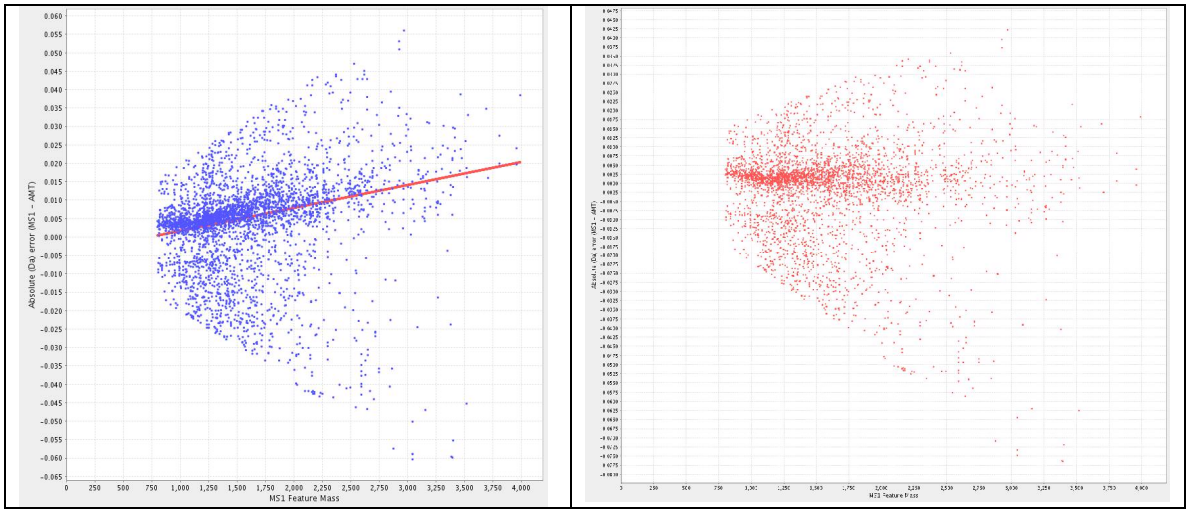
8. Augment MS/MS Peptide IDs with AMT Matches



9. Infer Proteins Across Fractions

Combine MS1 and MS/MS Data within Each Fraction

Protein Inference



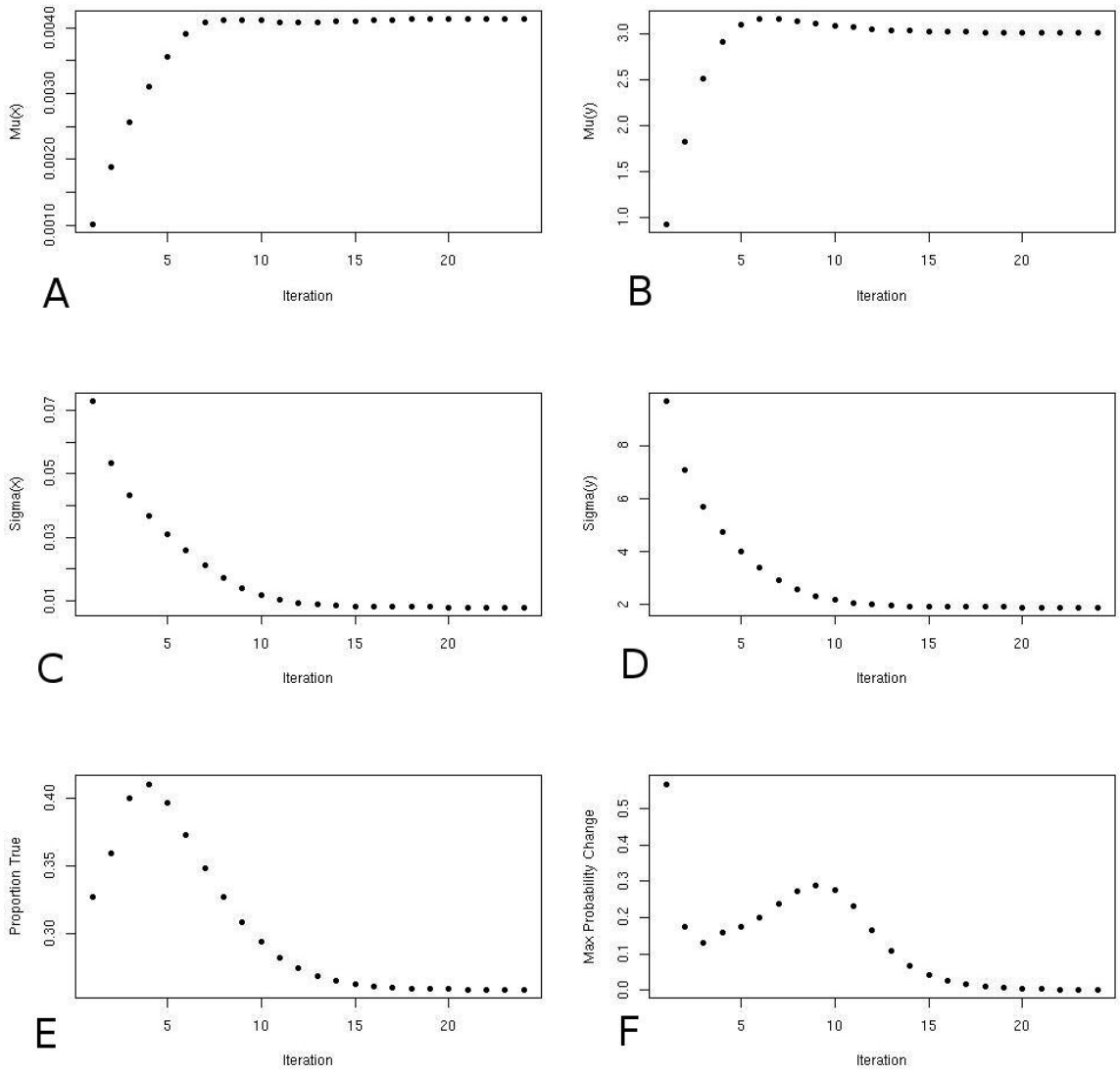
Methods

As mentioned in the manuscript, proper calibration of peptide masses is essential for the performance of AMT methods in general, and our probability calculation in particular. The following charts provide a visualization of LC-MS peptide feature mass calibration, before and after the recalibration step described in the manuscript. Mass miscalibration is often a linear function of peptide mass, so in both of these charts the X axis represents peptide mass. Each point represents a single AMT match with loose tolerances (e.g., the red points in the chart above). The Y axis represents deltaMass, i.e., the difference between LC-MS feature mass and AMT peptide mass. The line on the first chart represents the result of a robust linear regression using the two variables. This regression result is used to calibrate the LC-MS feature masses. The second chart shows the same data post-calibration.

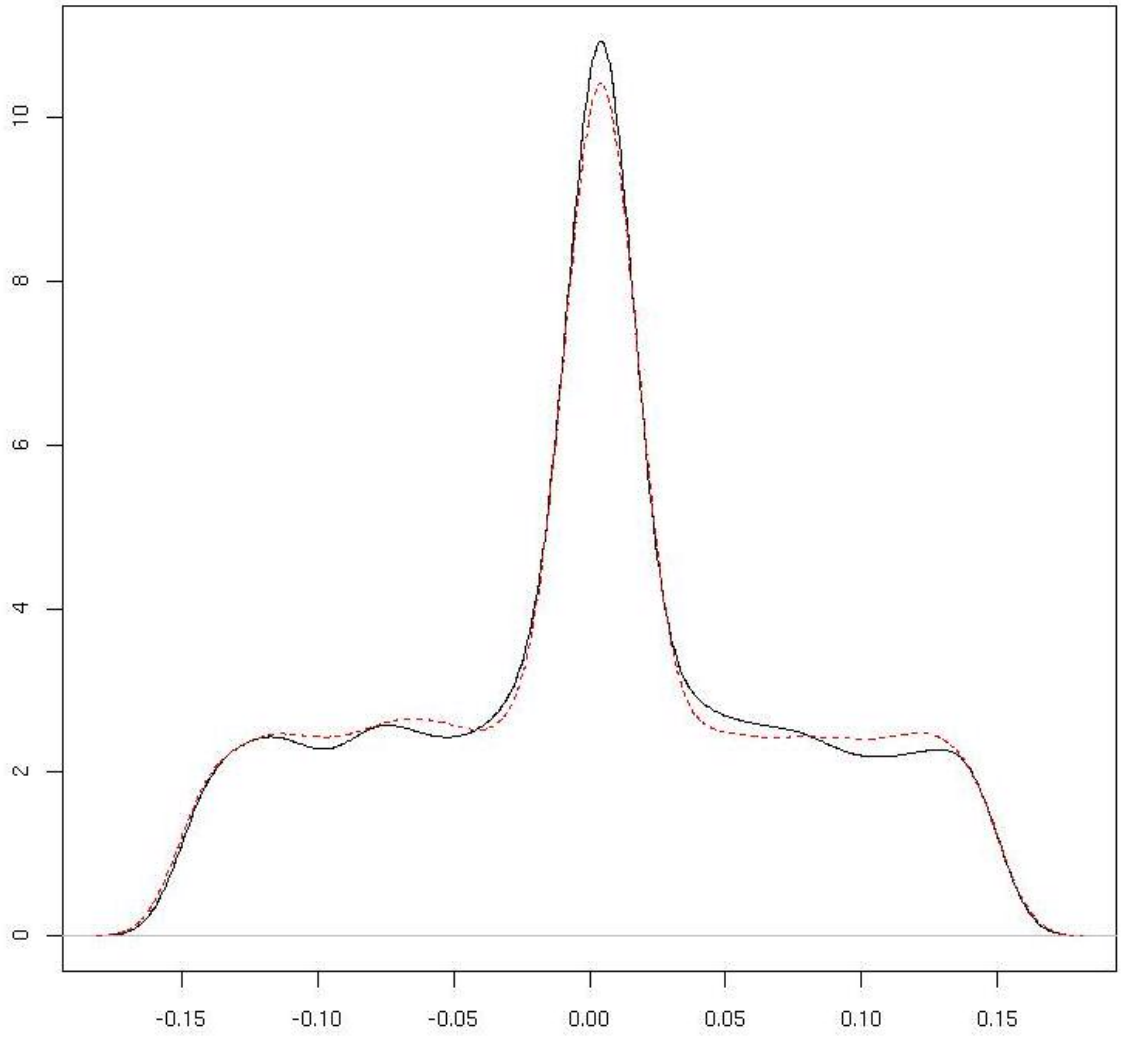
Results

The manuscript references several charts used for visual confirmation of the performance of the Expectation-Maximization (EM) algorithm used to estimate the parameters of the true and false match distributions. All of these charts are generated automatically by msInspect as part of the probability-estimation process.

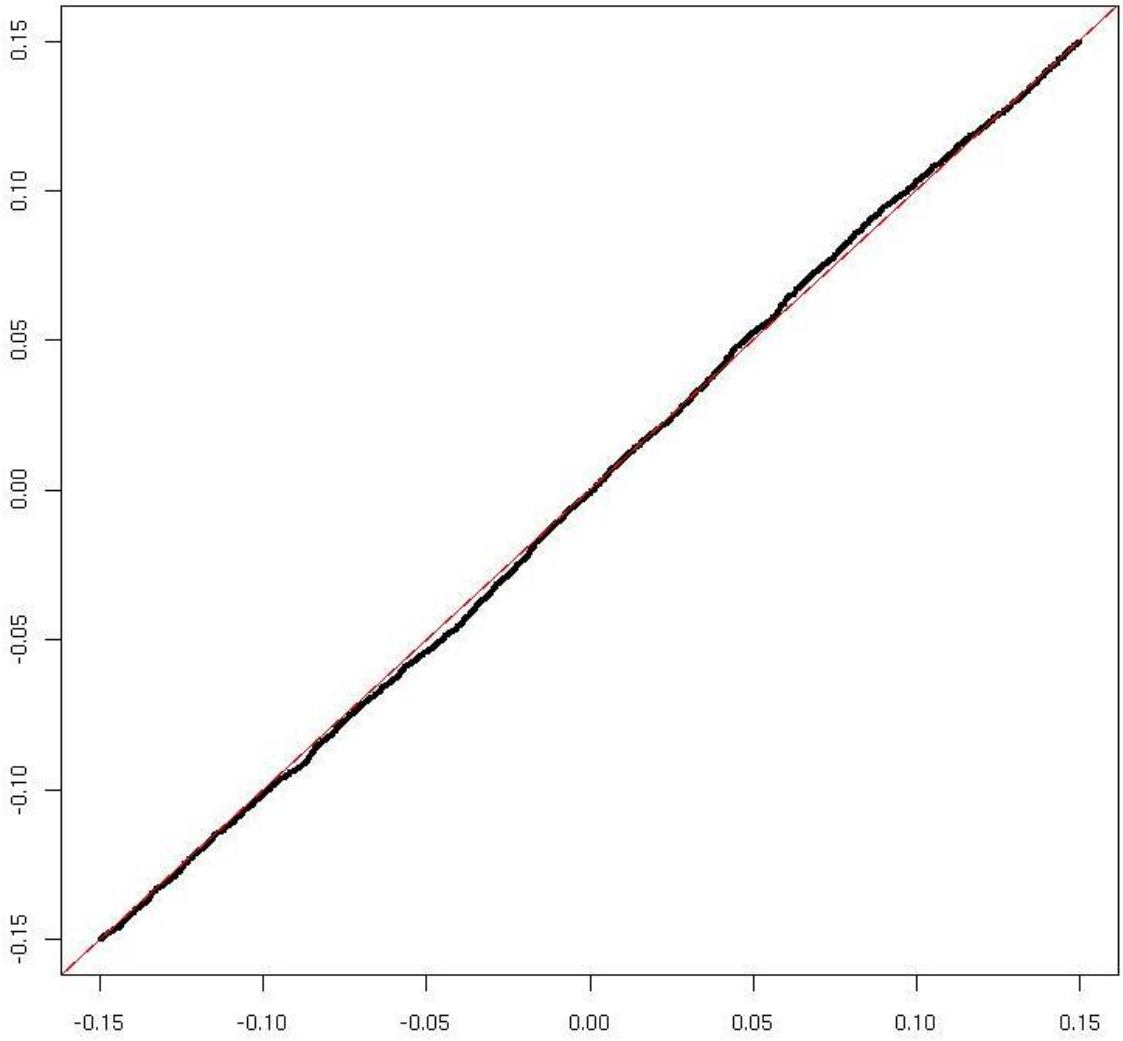
This first chart demonstrates the convergence of the various parameters of the two distributions (μ_{MASS}, σ_{MASS}, μ_{NRT}, σ_{NRT}, and the proportion of the number of points in one distribution vs. the other) over a number of iterations of the EM algorithm.



The next chart shows the density of the actual error data (solid line), in the Mass dimension, overlaid with the density of data generated from the estimated mixed distribution (dashed line). Perfect model performance would be indicated by indistinguishable distribution densities.



The following chart is a quantile-quantile plot of the actual error data (x axis), in the Mass dimension, vs. the data generated from the estimated mixed distribution. Perfect model performance would be indicated by a 1:1 line (shown in red).



Finally, this next chart combines the two dimensions of the actual data density (gray) and the estimated mixed distribution density (red) to visualize the agreement of the model with the actual data. A well-performing model overlays the actual data density very closely.