mspms: a R package and graphical interface for the processing and analysis of multiplex substrate profiling by mass spectrometry for proteases (MSP-MS) data.

**Abstract:**

Multiplex Substrate Profiling by Mass Spectrometry (MSP-MS) is a powerful method used to determine the substrate specificity of proteases. This method is of interest for many groups interested in the study of proteases and their role as regulators of many biological pathways whether applied to the study of disease states, the development of diagnostic and prognostic tests, generation of tool compounds, or rational design of protease targeting therapeutics. Analysis of the MS based data produced by MSP-MS is a multistep process involving detection and quantification of peptides, normalization, outlier detection, imputation, and cleavage sequence identification. This process can be challenging, especially for biologists/ mass spectrometrists with limited programming experience. To overcome these issues, we provide the mspms R package alongside a companion graphic user interface hosted at https://gonzalezlab.shinyapps.io/mspms\_shiny/ to facilitate the analysis of MSP-MS data utilizing good software/data analysis practices.

Introduction:

**Describe the method.**

Multiplex substrate profiling by mass spectrometry for proteases (MSP-MS) is a method for determining protease substrate specificity. This method works by utilizing a rationally designed library of peptide sequences, incubating them in a protease/protease mixture, potentially treating with an experimental condition, and subsequently detecting and quantifying the cleaved peptides through mass spectrometry.

**What is the problem?**

MSP-MS is applicable to a wide group of researchers as it only requires a synthetically created peptide library, mass spectrometer, and samples containing protease(s) to perform. While relatively simple conceptually, MSP-MS produces high dimensional data that is challenging for traditionally trained biologists or mass spectrometrists to analyze. Adequate interpretation of the MSP-MS data requires several steps. The data must be median normalized, outliers detected and removed, missing values imputed, and cleavage motifs must be recognized.

A central component of any experimental method working on complex data, is the data analysis pipeline. Poorly documented analysis code poses a number of challenges. It has been a prevalent problem in the biological research world, leading to the retraction of several high-profile papers in recent years. There is also a logistical problem, as it limits the portability/ reproducibility of analysis lab to lab since it requires specific knowledge from the programmer to successfully utilize. This is commonly a problem in academic labs, where a researcher may develop code widely used for analysis within the lab before leaving for a new opportunity. Often, this results in an opaque codebase that may be treated as a black box and can ultimately result in erroneous data analysis, especially as the original code gets distorted over time through decentralized use by various researchers. Solutions for this exist in the domain of software engineering, but they are infrequently successfully implemented in biology labs, due to the time it would take to develop and computer science knowledge required.

**How does this fix the problem?**

Here we describe mspms, which is a well-documented, portable, and reproducible R package for the analysis of MSP-MS data. This package brings good software practices to MSP-MS data analysis and offers a widely applicable data analysis tool for the protease research community.

There is also a wide community of protease researchers who may be interested in the MSP-MS method but find the R programming environment to be inaccessible. For such cases, we provide mspms-shiny, a graphical user interface to the core features of the mspms package. Mspms-shiny is available on the web at https://gonzalezlab.shinyapps.io/mspms\_shiny/, or by downloading and running locally on the researcher’s computer of choice.

**What will we do in this study to show that it works?**

Here we describe the core features of the package and demonstrate its use through the analysis of a novel dataset.

Methods:

**Package building, where is it hosted.**

Mspms was implemented in R and released under a MIT license.

**Preprocessing**

Mspms is designed to be downstream of computationally intensive solutions for detecting and quantifying peptides. Currently, it supports output from PEAKS, but additional file converters from popular MS are in development. Proteome Discoverer , FragPipe, MaxQuant.

**Data Normalization**

Data is

**Outlier Removal**

Data is

**Data Imputation**

Data is

**Cleavage Motif**

Data is

**Convenience Functions**

A number of functions are provided to provide convenient downstream analysis of the processed data. Pairwise t-tests are performed for each condition per timepoint. Anova is performed.

Plotting function for showing time course of peptide intensity.

High level representation of data via PCA plots or heatmaps with heirchical clustering performed.

**Code Availability**

This package is on CRAN?? Or Bioconducter?

The development version of mspms is available on github at <https://github.com/baynec2/mspms>.

The shiny interface to mspms can be found hosted at https://gonzalezlab.shinyapps.io/mspms\_shiny/

Results:

To illustrate how mspms analyzes, we analyzed a

Discussion:

We found that …. Is

mspms is a simple-to-use tool, that is specifically designed for biologists or mass spectrometrists who wish to analyze data produced through the MSP-MS method.