## Supplementary Data

# Title: Statistical modeling for enhancing discovery power of citrullination from tandem mass spectrometry data

Sunghyun Huh[†], Daehee Hwang[*‡] and Min-Sik Kim[*†]

[†]Department of New Biology, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu 42988, Republic of Korea

[‡]School of Biological Sciences, Seoul National University, Seoul 88026, Republic of Korea

*Correspondence:

Daehee Hwang, Ph.D.

Address: School of Biological Sciences, Seoul National University, Seoul 88026, Republic of Korea

E-mail: daehee@snu.ac.kr; Tel: +82-2-880-7522

Min-Sik Kim, Ph.D.

Address: Department of New Biology, DGIST, Daegu 42988, Republic of Korea

E-mail: mkim@dgist.ac.kr; Tel: +82-53-785-1630

## Table of Contents

## Supplementary Figures and Legends

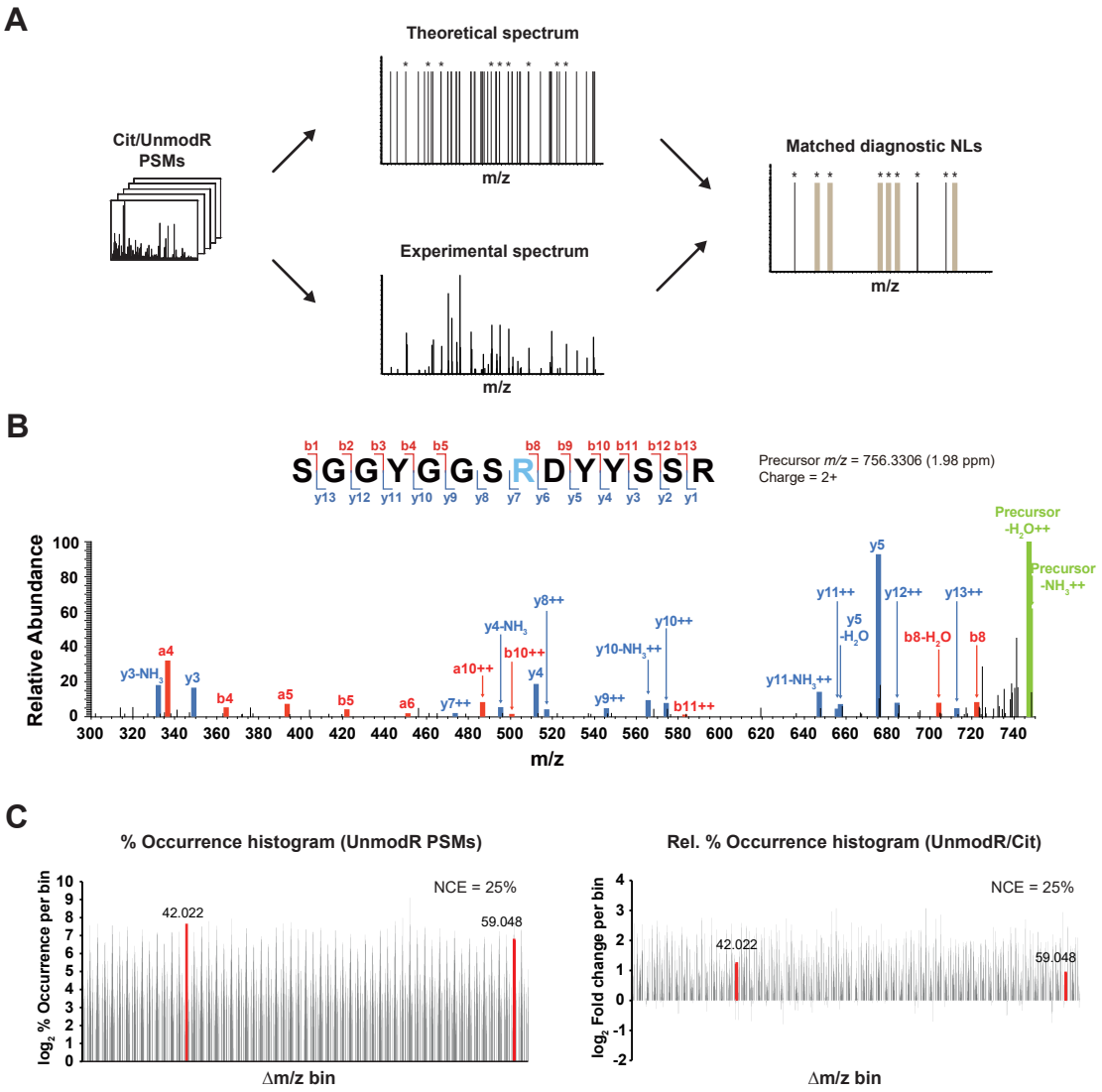


**Figure S1**. **Diagnostic neutral losses (NLs) for citrullination**. (A) Schematic workflow of theoretical spectrum generation for each peptide (left) and identification of matches (right, gray peaks; top, asterisk) by comparing the theoretical spectrum (top) with the corresponding experimental tandem-mass spectrometry (MS/MS) spectrum (bottom). (B) MS/MS spectrum of a representative unmodified (UnmodR) peptide spectrum match (PSM) showing the absence of diagnostic HNCO NLs. *y*- and *b*-ions and precursor ions are indicated as blue, red, and green lines, respectively. (C) Distributions of % occurrence in UnmodR PSMs and relative % occurrences (% occurrence-fold-changes of UnmodR/Cit). NLs from Arg producing 42.0218 and 59.0483 Da neutral fragments are indicated as red lines.

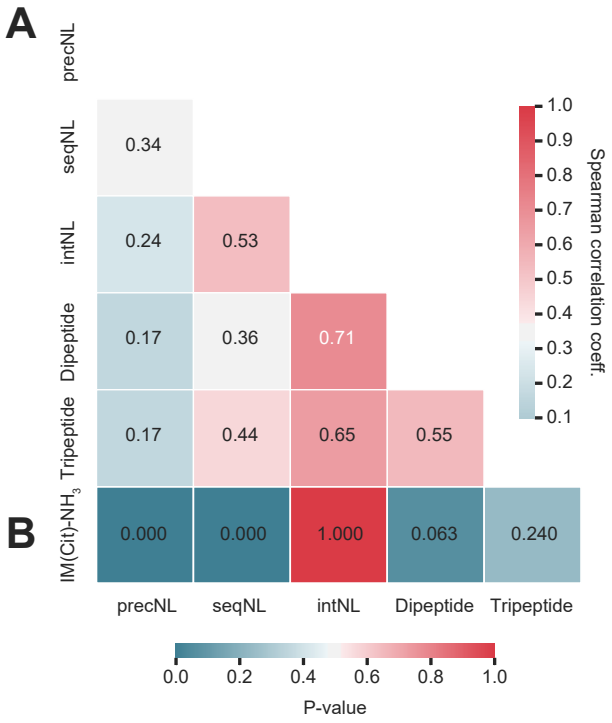


**Figure S2. High correlations in the occurrence of the diagnostic ions.** (A) Heat map depicting the Spearman correlation coefficients for pairwise comparisons of the five diagnostic ions- dipeptide, tripeptide, precursor neutral losses (precNL), sequence ion neutral losses (seqNL), and internal ion neutral losses (intNL). For each pair of the diagnostic ions indicated, Spearman correlation coefficients were computed using the numbers of the occurrences of the diagnostic ions across PSMs. For a correlation coefficient, a P-value was estimated using an empirical t-test (see below). (B) Heat map showing the significance (P-value) of the co-occurrence of immonium ion of Cit with loss of $NH_3$ [IM(Cit)-$NH_3$] with the other five diagnostic ions. The occurrence of IM(Cit)-$NH_3$ is the binary value (i.e., zero when not observed in a PSM while one when observed), unlike the other diagnostic ions with the discrete values (i.e., the numbers of the occurrences in a PSM). We first transformed the discrete values to the binary values by changing the non-zero values into ones and then estimated the significance (P-value) in the co-occurrence of IM(Cit)-$NH_3$ and each diagnostic ion (i.e., non-zero values for IM(Cit)-$NH_3$ and the diagnostic ion) using an empirical Fisher's exact test. For the empirical t-test (or Fisher's exact test), we randomly permuted the discrete (or binary) values 10,000 times, estimated an empirical null distribution of the t-statistic (or Fisher's statistic) from t-statistic (or Fisher's statistic) values for the randomly permuted values, and finally calculated P-values by applying the right-sided tests to the observed t-statistic values for Spearman correlation coefficients (or Fisher's statistic for the co-occurrences) using the estimated empirical null distribution.

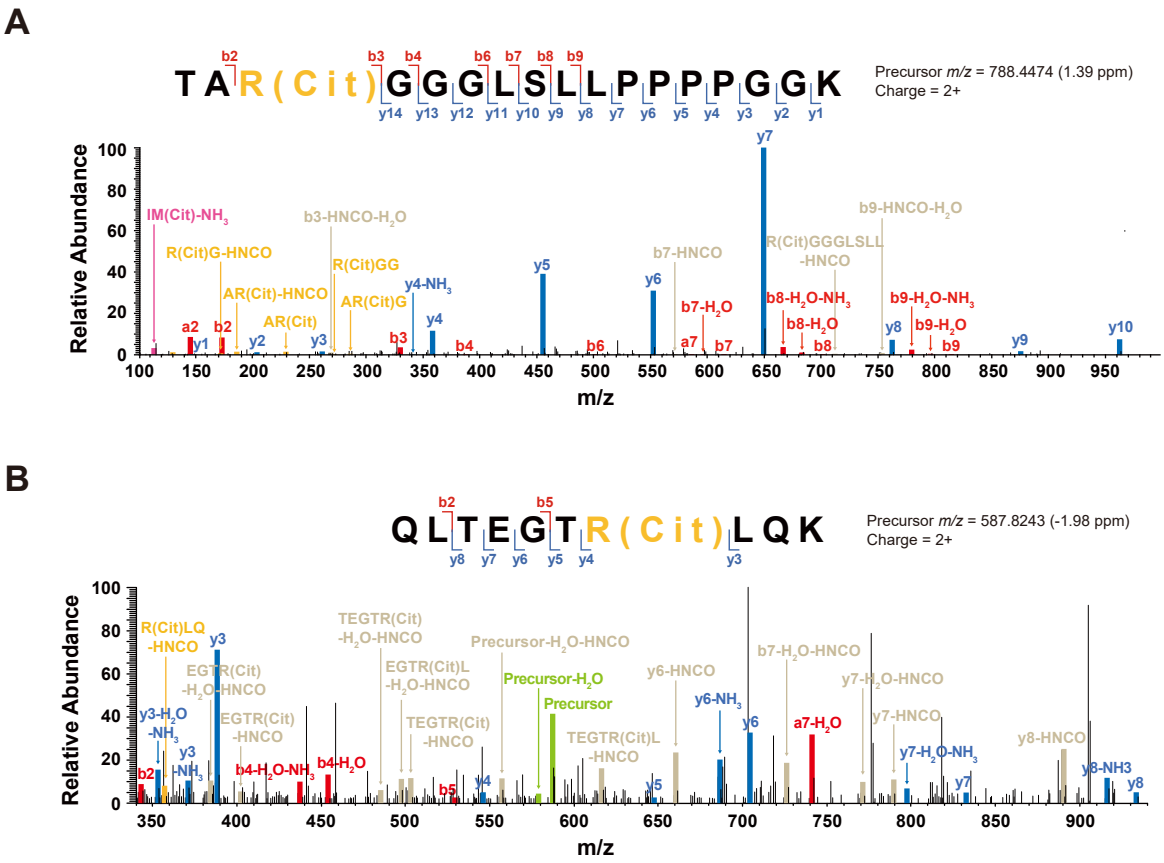


**Figure S3**. **MS/MS spectra of representative citrullinated (Cit) PSMs newly predicted by EN model.** (A-B) Annotated MS/MS spectra of Cit PSMs from the proteome of the synthetic peptides (A) and the brain proteome (B) predicted by the elastic net (EN) logistic regression model**.** Diagnostic IM(Cit)-$NH_3$, internal ions (INTs), and NLs are indicated as magenta, yellow and gray arrows, respectively.

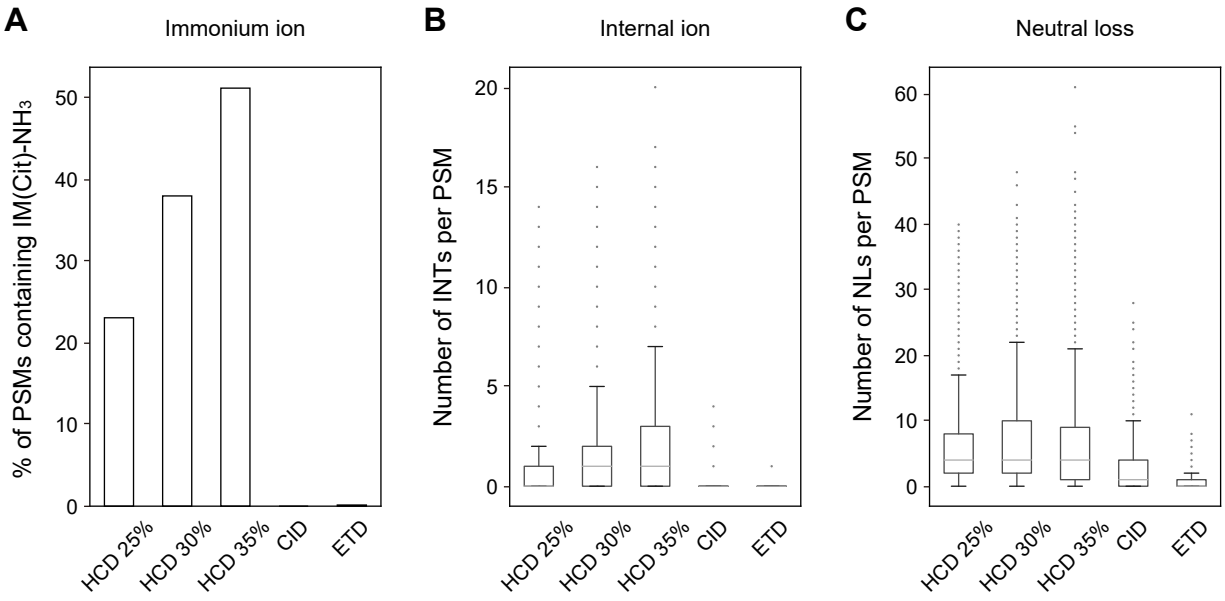


**Figure S4. Dependency of production of diagnostic ions for citrullination on MS instrument characteristics.** The production of diagnostic ions including (A) IM(Cit)-$NH_3$, (B) INTs, and (C) NLs was compared in the data generated by HCD with three different collision energies (20, 30, and 35%), CID, and ETD. Percentage of PSMs for IM(Cit)-$NH_3$ and the numbers of INTs and NLs per PSM are displayed.

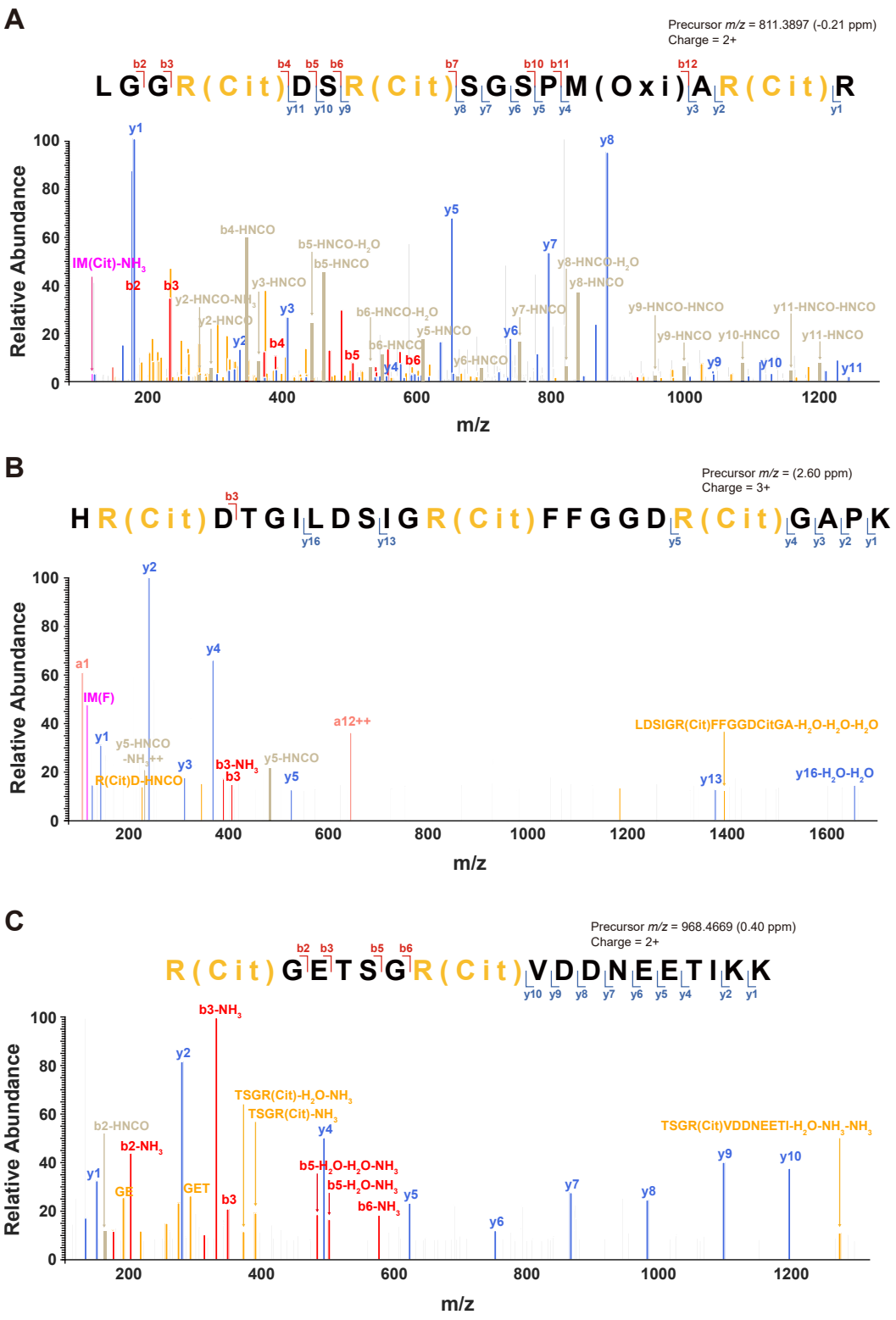


**Figure S5. MS/MS spectra of representative multiply Cit PSMs.** MS/MS spectra of a true positive multiply Cit PSM (A), a false positive multiply Cit PSM (B), and a false negative multiply Cit PSM (C). IM(Cit)-$NH_3$, ITNs, and NLs are indicated as magenta, yellow and gray lines, respectively.

| $\Delta m/z$ | Identity | Reference |
|---|---|---|
| 12.9896 | $\Delta m/z$ of two NLs (HNCO & $NH_3$) in the same doubly charged ions | Zolg et al.[22] |
| 21.5029 | NL of HNCO in doubly charged ions | |
| 25.9793 | $\Delta m/z$ of two NLs (HNCO & $NH_3$) in the same singly charged ions | |
| 43.0058 | NL of HNCO in singly charged ions | |

**Table S1. Four diagnostic NL candidates.** $\Delta m/z$ values and identities are shown with the literature that previously reported them.