An azo coupling-based chemoproteomic approach to systematically profile the tyrosine reactivity in the human proteome

Fangxu Sun, Suttipong Suttapitugsakul, and Ronghu Wu*

[1]School of Chemistry and Biochemistry and the Petit Institute for Bioengineering and Bioscience, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

ABSTRACT

The tyrosine residue of proteins participates in a wide range of activities including enzymatic catalysis, protein-protein interaction, and protein-ligand binding. However, the functional annotation of the tyrosine residues on a large scale is still very challenging. Here, we report a novel method integrating azo coupling, bioorthogonal chemistry, and multiplexed proteomics to globally investigate the tyrosine reactivity in the human proteome. Based on the azo-coupling reaction between aryl diazonium salt and the tyrosine residue, the two different probes were evaluated, and the probe with the best performance was employed to specifically target the tyrosine residues. After the reaction, tagged tyrosine containing peptides were selectively enriched using bioorthogonal chemistry, and a small tag on the peptides from the cleavage perfectly fits for site-specific analysis by MS. Coupling with multiplexed proteomics, we quantified over 5,000 tyrosine sites in MCF7 cells and these quantified sites displayed a wide range of reactivity. The tyrosine residues with high reactivity were found on functionally and structurally diverse proteins, including those with the catalytic activity and binding property. This method can be extensively applied to advance our understanding of protein functions and facilitate the development of covalent drugs to regulate protein activity.

INTRODUCTION

With the rapid development of mass spectrometry (MS)-based proteomics,[1-3] large-scale characterization and quantification of proteins have significantly advanced in recent years and provided a wealth of information including protein expression, spatial distribution, dynamics, and interactions.[4-6] Nevertheless, the biological functions of many proteins in the eukaryotic and prokaryotic organisms remain elusive. Comprehensive analysis of protein post-translational modifications (PTMs), which refers to the covalent modifications of proteins, including glycosylation, phosphorylation, and ubiquitination, has attracted much attention,[7-15] and is effective to investigate the biological functions of proteins on a system-wide level. In addition, the amino acid residues of proteins create a specific microenvironment and have critical roles in determining the functional activities of proteins. Correspondingly, the same amino acid residue may have different reactivities in different local environments. Previously, global analysis of the reactivity of the cysteine residues of proteins was achieved by integrating a covalent probe (iodoacetamide-alkyne) and a quantitative proteomics approach. The authors found that the cysteine residues with hyper-reactivity were involved in different types of functional activities, including enzymatic catalysis and redox regulation, which greatly expand our understanding of protein functions.[16]

Besides revealing their biological functions, large-scale analysis of the reactivity of the amino acid residues also facilitates the discovery of proteins as drug targets that can be targeted by covalent ligands. Although many proteins have been reported to be correlated with various diseases, it is still challenging to develop drugs to specifically target them and some are even considered as undruggable. The development of covalent ligands offers another strategy to expand the landscape of proteins amenable to be targeted by small molecules. After investigating

the reactivity of the amino acid residues using MS-based proteomics, a specific electrophilic reagent can be designed to target certain sites. For example, reactive amino acid residues are often found in the binding pockets of enzymes, which can be targeted by covalent inhibitors,[17-18] indicating that comprehensive profiling of the reactivity of the amino acid residues using MS-based proteomics promotes the development of covalent drug inhibitors.

Besides cysteine, other amino acid residues have also recently been studied using MS-based proteomics. The lysine residues are frequently labeled with N-hydroxysuccinimide-esters (NHS-esters), and recently over 9,000 lysine resides were profiled in human cells with a STP (sulfotetrafluorophenyl) ester-based probe.[19] Furthermore, the aspartate and glutamate residues in human cells and bacteria were investigated using a 2H-azirine-based probe and a light-activatable 2,5-disubstituted tetrazoles-based probe, respectively.[20-21] A method called redox activated chemical tagging (ReACT) was developed to specifically target methionine in biological systems and applied for chemoproteomic identification of functional methionine residues, which found a group of proteins with hyperactive methionine residues including enzymes, chaperones, and nucleoproteins, and structural proteins.[22]

The tyrosine side chain plays a critical role in a variety of biological functions of proteins because of its uniquely structural and electronic features, and has been reported to be selectively targeted by covalent inhibitors under specific conditions.[23-25] For example, Chen and coworkers were able to specifically modify a lipid-binding protein, i.e. cellular retinoic acid binding protein 2 (CRABP2), using aryl fluorosulfates. The aryl fluorosulfate warhead with the low reactivity was used to covalently target tyrosine through the proximity effect.[26] Although the tyrosine residue is involved in various biological processes and regulates different protein functions, global analysis of tyrosine still lags behind. A systematic study of the tyrosine residues will not

only reveal their biological functions, but also help identify proteins that can be covalently targeted by small-molecule ligands. Recently, Hahm et al. globally characterized the tyrosine residues using sulfur-triazole exchange (SuTEx) chemistry and studied tyrosine phosphorylation changes with pervanadate activation.[27-28] Effective methods to analyze tyrosine on proteins will further advance our understanding of the tyrosine residues, which not only reveals their novel biological functions, but also expands the scope of covalent drug candidates.

In this work, we developed a novel and effective method integrating azo coupling, bioorthogonal chemistry, and multiplexed quantitative proteomics to globally investigate the tyrosine residues in the human proteome. The azo-coupling reaction between aryl diazonium salt and the aromatic amino acid residues, such as tyrosine, was reported to modify the protein surface.[29-33] Through changing the substituent groups on the phenyl ring of the aryl diazonium, its reactivity towards tyrosine can be tuned.[34] Reactive tyrosine normally possesses a high level of nucleophilicity on the oxygen atom and tends to display higher electron density, which also enhances the nucleophilicity of the phenyl ring on tyrosine. Therefore, an aryl diazonium probe that specifically targets the tyrosine residues through the electrophilic-aromatic substitution process can be used to study reactivities of tyrosine. Combining azo coupling with multiplexed quantitative proteomics, we achieved systematic and site-specific analysis of the tyrosine residues in human cells. This method can be extensively applied to study the tyrosine residues in the biological and biomedicine research fields.

EXPERIMENTAL SECTION

Cell culture

MCF7 cells (ATCC) were cultured in Dulbecco's Modified Eagle's Medium (DMEM, Sigma-Aldrich) with high glucose supplemented with 10% fetal bovine serum (FBS, Corning) and 1% penicillin-streptomycin solution (Sigma-Aldrich). When the confluency reached ~90%, the cells were washed with PBS three times, harvested by scraping, and pelleted by centrifugation (300 *g*, 5 min). The cells were further washed with ice-cold PBS.

Tyrosine labeling, click chemistry, and protein digestion

The diazonium-alkyne probe was synthesized according to previous reports.[35-36] Briefly, *m*-ethynylaniline or *p*-ethynylaniline (Sigma-Aldrich) was mixed with 1 M HCl on ice for 15 min. A freshly prepared sodium nitrite ($NaNO_2$, Sigma-Aldrich) aqueous solution was chilled on ice for 15 min and added slowly to the ethylaniline solution above. The mixture was incubated on ice for 45 min. Subsequently, a tetrafluoroboric acid solution (HBF_4, 48%, Sigma-Aldrich) was added to precipitate the diazonium-alkyne probes. After filtration, the product was purified through being dissolved in acetonitrile (ACN) and then precipitated by adding cold ether (Sigma-Aldrich). The purification procedure was repeated one more time.

The cell pellets were resuspended in PBS and lysed by rapid freeze-thawing three times with liquid nitrogen. After centrifugation (5000 *g*, 10 min), the supernatant was transferred to a new tube. The diazonium-alkyne probe **1** was added to the cell lysates to a final concentration of 50 μM (low concentration, L) or 500 μM (high concentration, H), respectively. The labeling reaction lasted for 60 min at room temperature (RT) with end-over-end rotation. The probe was removed, and proteins were isolated through the methanol-chloroform precipitation method. The purified proteins were solubilized in PBS with 0.4% SDS *via* sonication. The biotin-alkyne (25.0

mM stock solution in DMSO, Click Chemistry Tools) was added to the solution to a final concentration of 250 μM, followed by $CuSO_4$ and tris(3-hydroxypropyltriazolylmethyl) amine (THPTA, Click Chemistry Tools) to the concentrations of 1.0 and 5.0 mM, respectively.[37] Sodium ascorbate and guanidine hydrochloride were transferred to the solution at 15.0 mM to initiate the copper(I)-catalyzed azide alkyne cycloaddition (CuAAC) reaction. The reaction was incubated at RT for 2 h. Eventually, the proteins were purified again, and then digested with trypsin (Promega) in a digestion buffer (50 mM HEPES pH 8.6, 1.5 M urea) at 37 ºC overnight. Trifluoroacetic acid (TFA, Fisher Scientific) was added to quench the digestion reaction by adjusting the pH value to be <2. The peptides were desalted using a tC18 Sep-Pak cartridge (Waters) and dried under vacuum.

Enrichment of labeled tyrosine peptides

The dried peptides were dissolved in PBS and incubated with NeutrAvidin beads (Fisher Scientific) at RT for 2 h. The NeutrAvidin beads were transferred to a spin column (Fisher Scientific) and washed with PBS for ten times. Then, the tyrosine-modified peptides were eluted from the beads by incubating with a freshly prepared 25.0 mM sodium dithionite ($Na_2S_2O_4$) solution for 30 min. The elution step was repeated one more time and the elutes were combined. The enriched peptides were purified using a tC18 Sep-Pak cartridge.

TMT labeling and peptide fractionation

The peptides from six samples were labeled with the six-plex tandem mass tag (TMT) reagents (Thermo Scientific) according to the manufacturer's protocol. Channels 126, 127, and 128 were used to label the peptides from the samples reacted with the probe at the low concentration, and channels 129, 130, and 131 were for the peptides reacted at the high concentration. Briefly, the peptides from each sample were dissolved in the HEPES buffer (pH 8.6, 200 mM, 100 μL). Then ACN (30 μL) was added to each sample above. Each tube of the TMT reagent was warmed to RT and then dissolved in ACN (41 μL), and 5 μL was used for the peptide labeling. The TMT labeling reaction lasted for 60 min at RT with shaking, and subsequently was quenched with hydroxylamine (10 μL, 5%, Sigma-Aldrich) for 15 min. The labeled peptides from each sample were mixed, purified, and dried. The peptides were dissolved in 300 μL ammonium acetate (pH 10, 10 mM) and then loaded onto a reversed-phase C18 HPLC column (Waters). The peptides were separated into 16 fractions with a 40-min gradient of 5−55% ACN containing ammonium acetate (pH 10, 10 mM). Each fraction was purified using the StageTip method before LC-MS/MS analysis.

LC-MS/MS analysis

The dried, TMT-labeled peptides were dissolved in the loading buffer (5% ACN and 4% formic acid (FA)), and 4 μl was loaded onto a reversed-phase microcapillary column packed with C18 beads through a WPS-3000TPLRS autosampler (UltiMate 3000). The peptides were first separated by HPLC *via* an UltiMate 3000 HPLC system, followed by being detected in a hybrid dual-cell quadrupole linear ion trap-Orbitrap mass spectrometer (LTQ Orbitrap Elite, Thermo Scientific). A data-dependent Top15 method was used for peptide detection. A full MS scan

(resolution: 60,000) was recorded in the Orbitrap at the automatic gain control (AGC) of 10^6. The top 15 precursor ions with the highest intensities were selected for fragmentation using higher-energy collision dissociation (HCD) and the normalized collision energy (NCE) was set to 40%. The fragments were then detected in the Orbitrap cell with high resolution and high mass accuracy. The selected precursor ions were excluded for 90s. Ions with a single or unassigned charge were not fragmented.

Database searching, data filtering, and bioinformatic analysis

The SEQUEST algorithm (version 28)[38] was used to search the raw files against the database containing all human proteins downloaded from UniProt (*Homo sapiens*). The search parameters were set as indicated below: precursor mass tolerance (10 ppm); product ion mass tolerance (0.025 Da); digestion enzyme (trypsin); missed cleavages (up to three). The fixed modifications included carbamidomethylation of cysteine (+57.0214), TMT tag of lysine and the peptide N-terminus (+229.1629), and the variable modifications included oxidation of methionine (+15.9949), tyrosine (Y) with the modified tag (+15.0109). The target-decoy method was employed to evaluate the false discovery rates (FDRs) for peptide and protein identifications.[39] Linear discriminant analysis (LDA) that considers multiple parameters (XCorr, precursor mass accuracy, and charge state) was performed to control the degree of accuracy of probe-modified peptide identifications,[40] and peptides with <7 amino acid residues were removed. The FDRs of probe-modified peptides and proteins were both controlled to <1%. The ModScore was used to evaluate the accuracy of the site localization, and sites with a ModScore >13 ($P < 0.05$) were

considered to be well-localized.[41] The intensities of the TMT reporter ions in the tandem MS were used to quantify the identified peptides.

The tyrosine phosphorylation sites were downloaded from PhosphoSitePlus (HTP score ≥ 1).[42] The amino acid frequency was generated with pLogo.[43] The protein-protein interactions (PPIs) were extracted from String and visualized by Cytoscape.[44-45] Functional analysis of the protein interaction network was performed using String and PANTHER (Protein Analysis Through Evolutionary Relationships).[46] ClusterMaker in Cytoscape was exploited to perform protein cluster analysis with the MCL (Markov Clustering) algorithm.[47] The inflation parameter was set to 3. The protein clusters were also compared with the protein complexes deposited in CORUM.[48] Protein domains were extracted from SUPFAM.[49] The crystal structures were downloaded from RCSB Protein Data Bank and visualized using PyMOL.[50] For ENPP1, the crystal structure (PDB: 4B56) of its analog from mouse was used.

Gel-based fluorescence assay

As described above, proteins reacted with the diazonium-alkyne probe at different concentrations for 60 min after cell lysis. The diazonium-alkyne probe was removed by passing the cell lysates through a Bio-Gel P-6 column (Bio-Rad). Then, azide-fluor 545 (Sigma-Aldrich) was added to the cell lysates (the final concentration of 100 μM), followed by $CuSO_4$, THPTA, and sodium ascorbate. The reaction lasted for 60 min with shaking at RT. The SDS-PAGE loading buffer without DTT (DTT tends to reduce the formed diazo group during the SDS-PAGE) was added, and then proteins were separated by SDS-PAGE. The gel was further stained with SimplyBlue.

The fluor545 signal and SimplyBlue staining bands were visualized and recorded using a GE Typhoon Trio+ Fluorescence/Phospho-Imager system.

RESULTS AND DISCUSSION

Principle of large-scale analysis of the tyrosine residues in the human proteome

Comprehensive analysis of the tyrosine residues on proteins offers an effective and straightforward way to systematically investigate the potentially reactive tyrosine sites and reveals novel biological functions of corresponding proteins on a large scale. This will facilitate the development of small molecules to covalently modulate the activities of proteins and to target proteins previously known as undruggable. Until now, a wide range of small molecule-based probes have been developed to specifically target various amino acid residues. Compared with other common nucleophilic amino acids such as cysteine, lysine, and serine, the structure of tyrosine is unique because there is a conjugative effect between the benzene ring and the oxygen atom. Therefore, we reason that an electrophilic aromatic substitution that targets the tyrosine residues can reveal their nucleophilicity and investigate the reactivity of the tyrosine residues.

The tyrosine residue was used for chemo- and site-selective protein modification through the electrophilic aromatic substitution. The unique properties of the tyrosine residues allow for their selective modifications with reactive aryl diazonium salts *via* electrophilic aromatic substitution to generate diazo compounds, which is also called azo coupling. However, up to date, the azo coupling has not been exploited to systematically study the reactivity of the tyrosine residues. Here, we developed an aryl diazonium-alkyne probe to selectively target the tyrosine residues in the human proteome and established a novel method to comprehensively profile the

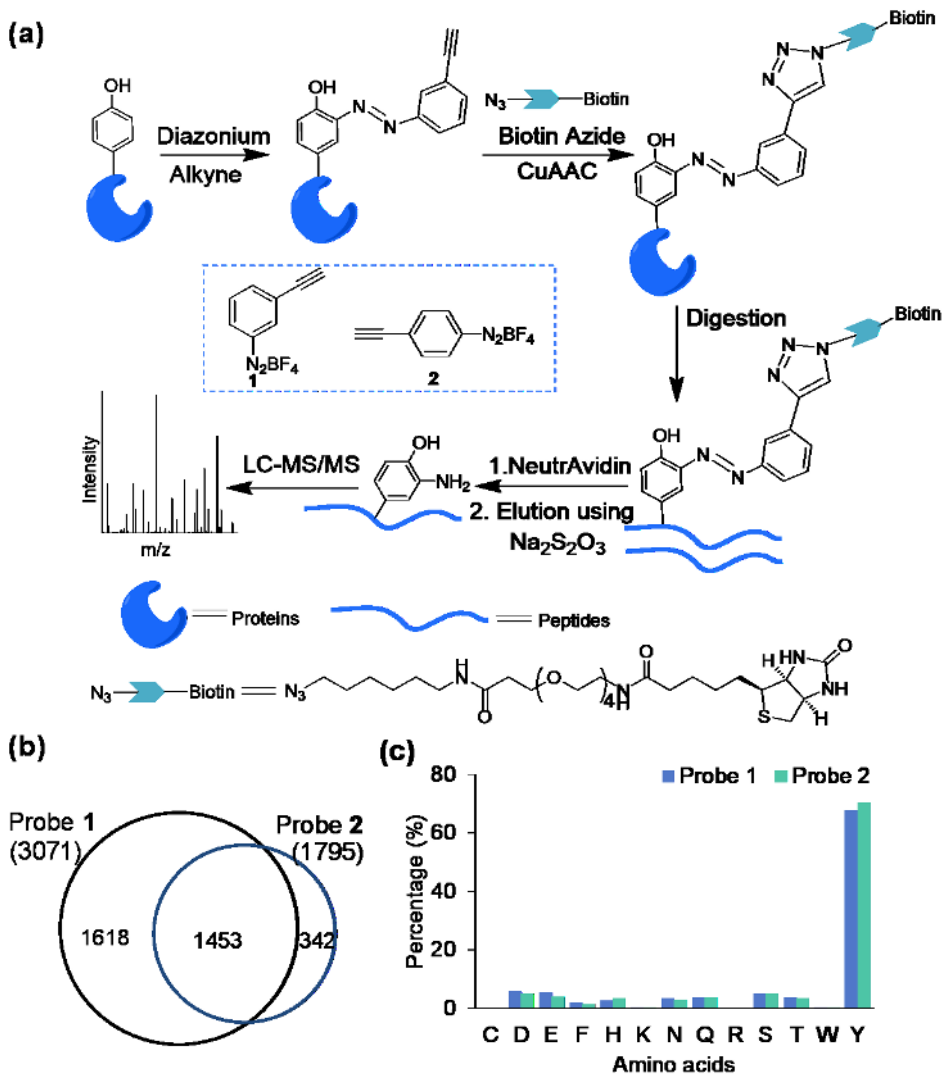

Figure 1. (a) Experimental procedure for comprehensive profiling of the tyrosine residues in the human proteome by integrating azo coupling and click chemistry (CuAAC). (b) Comparison of unique peptides with probe-modified tyrosine residues identified using probe **1** and probe **2**. (c) The experimental results demonstrate that both probes are specific to target the tyrosine residues.

tyrosine residues (Figure 1a). The reactivities of aryl diazonium salts can be tuned by changing the substituent groups. After the aryl diazonium-alkyne probe was conjugated to the tyrosine residue, a biotin tag was added to tyrosine-containing proteins through CuAAC. Then proteins

were digested, and peptides containing the biotinylated tyrosine were enriched using NeutrAvidin beads. The reaction between the side chain of tyrosine and the aryl diazonium-alkyne probe produced an azo group, which were cleaved by sodium dithionite. The remaining amino group on the modified peptides facilitates peptide ionization for MS analysis. The peptides containing the modified tyrosine can be site-specifically identified by MS.

Identification of the tyrosine residues with azo coupling and LC-MS/MS

To optimize the proteomics workflow for identifying the tyrosine residues in the human proteome, we first compared the protein-centric and peptide-centric enrichment strategies. For the peptide-centric approach, the workflow is described in Figure 1a. For the protein-centric approach (Figure S1a), the tagged proteins *via* azo coupling were enriched directly by the beads conjugated with the azide group through CuAAC, followed by on-bead digestion and elution with sodium dithionite. After LC-MS/MS analysis, the results showed that the peptide-centric method can identify many more peptides containing modified tyrosine compared to the protein-centric method (Figure S1b). The reason is that the enrichment at the peptide level can reduce non-specific binding and eliminate other peptides without the probe-modified tyrosine residues from proteins.

To further improve the coverage, we compared two aryl diazonium-alkyne probes (Figure 1a) with different reactivities for profiling the tyrosine residues. With the same experimental conditions, probe **1** and probe **2** were used to label the tyrosine residues, followed by biotinylation, protein digestion, peptide enrichment, and LC-MS/MS analysis. The results demonstrated that probe **1** outperformed probe **2**, and 71% more unique tyrosine-containing

peptides were identified using probe **1** compare to probe **2** (Figure 1b, Table S1). Although the structures of *m*-ethynylaniline (probe **1**)- and *p*-ethynylaniline (probe **2**)-based diazonium-alkyne probes are very similar, probe **1** possesses higher reactivity towards the aryl group because the alkyne group in the meta position is more electron-withdrawing, which can increase the reactivity of the corresponding diazonium salt. For both probes, most of the identified modified peptides are labeled at the tyrosine residues (Figure 1c), and with the fast speed of MS, the side reactions with other residues will not affect the quantification of the tyrosine reactivity. These results indicate that probe **1** is similarly specific as probe **2**, but more reactive for targeting the tyrosine residues in the human proteome. After finishing the experimental work, we found a recent study revealing that the most dominant modified sites came from cysteine instead of tyrosine for probe **2**.[51] This may be caused by the radical-based coupling between aryl diazonium salts and amino acid residues rather than azo coupling. Different sample preparation procedures, including the elution of the intact biotin tag in that report and the reduction of the azo group in this work, may be the reasons for the distinct labeling specificities. Compared to probe **2**, probe **1** has similar specificity towards the tyrosine residue, but it enabled us to identify many more tyrosine residues among the parallel experiments (3071 vs. 1795), as shown in Figure 1.

Quantification of the tyrosine residues in the human proteome

The concentration-dependent labeling strategy has been frequently used to evaluate the reactivity of amino acid residues including cysteine, lysine, aspartates, and glutamates, and reveal their potential biological functions. In this work, we employed probe **1** with better performance and quantitative proteomics to profile the tyrosine reactivity in the human proteome. The protein

labeling with probe **1** was concentration-dependent as evaluated by conjugating the labeled proteins with an azide dye using CuAAC and visualizing *via* in-gel fluorescence scanning (Figure S2), enabling us to study the reactivity of the tyrosine residues.

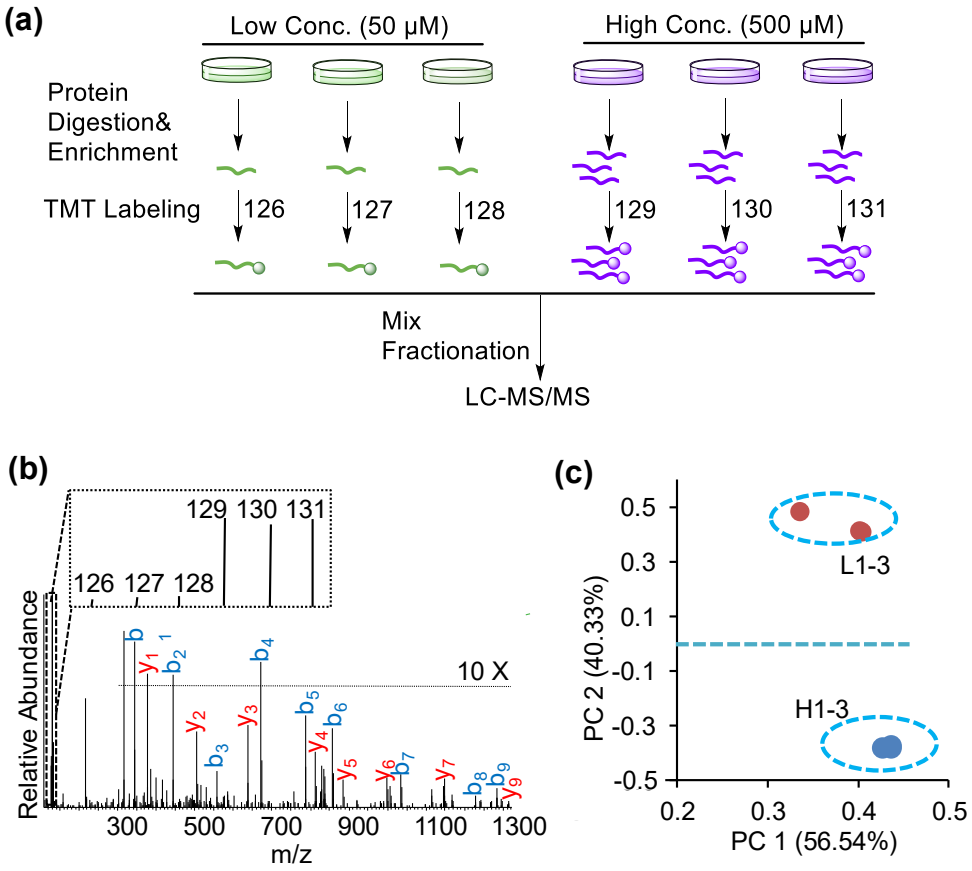

Figure 2. (a) Experimental procedure for quantifying the reactivity of the tyrosine residues by integrating azo-coupling and multiplexed proteomics. (b) An example tandem mass spectrum of the peptide ITLDNAY#MEK (# refers to the modified site). (c) PCA analysis of the peptide intensities of the samples treated with the high-concentration (H) and low-concentration (L) probe in the biological triplicate experiments.

We treated the whole cell lysates from MCF7 with high (500 μM) and low (50 μM) concentrations of probe **1**, respectively, followed by quantitative proteomics to evaluate the

reactivity of the tyrosine residues. Multiplexed quantitative proteomics with tandem mass tags (TMTs) as the labeling reagents has been widely used to quantify proteins because it can analyze many samples simultaneously, which reduces experimental time and increases the quantification accuracy. Here, we combined the azo labeling and multiplexed proteomics to evaluate the reactivity of the tyrosine residues in the biological triplicate experiments. The peptides labeled with each channel of the TMT reagents generate a unique reporter ion in the tandem MS, and the intensities of the reporter ions can be used to quantify peptides among the six samples (Figure 2a).

An example of peptide identification and quantification is shown in Figure 2b. The peptide of ITLDNAY#MEK was confidently identified with an XCorr of 4.5 and the mass accuracy of 3.6 ppm. The ModScore was determined to be 1000, indicating that modified site was localized on the tyrosine residue. The peptide is from PKM, an important glycolytic enzyme. The intensities of the reporter ions clearly demonstrated that the modification of Y148 was dependent on the probe concentration. To assess the similarity and difference in the high and low concentration-treated samples, we next performed principal component analysis (PCA). As shown in Figure 2c, all the replicates in the high concentration-treated samples clustered together and segregated from the three replicates in the low concentration-treated samples. The high Pearson correlation coefficient (>0.9) among the three replicates in each group further demonstrated the good reproducibility of the current quantification approach (Figure S3).

Over 5,000 tyrosine sites have been characterized from the experiments (Table S2). The majority of the sites (over 95%) are considered as well-localized with a ModScore larger than 13 ($P<0.05$), and about 93% have a ModScore larger than 19 ($P< 0.01$) (Figure 3a). Nearly half (48%) of the proteins contained only one modified tyrosine, while 18% identified proteins had at

least four modified tyrosine sites (Figure 3b). For example, glucose-6-phosphate 1-dehydrogenase (G6PD), which plays an important role in glycolysis, was detected with 13 tyrosine sites in this work. We also studied the relationship between the protein length and the number of the modified tyrosine residues per protein, but no significant correlation was found between them (Figure S4).

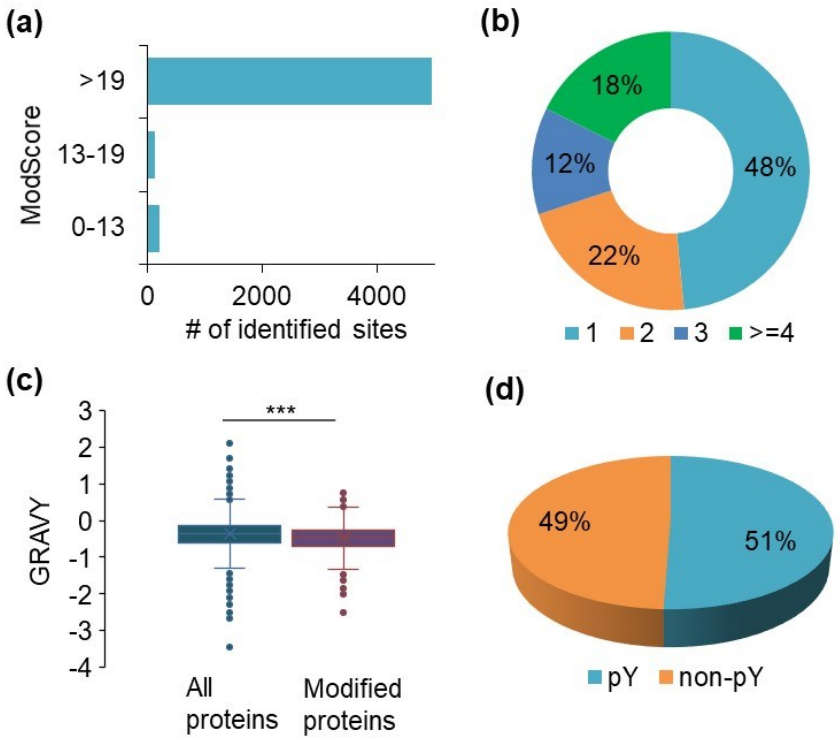

Figure 3. Systematic analysis of the tyrosine residues profiled in the quantification experiment. (a) The distribution of the ModScore values for the modified tyrosine sites. (b) The distribution of the number of the identified tyrosine residues per protein. (c) The GRAVY values of proteins with the modified tyrosine residues and the values from all proteins in the human proteome (***P<0.001). (d) The overlap between the identified tyrosine residues and the tyrosine phosphorylation sites in the PhosphoSitePlus database.

The GRAVY values, indicating the extent of hydrophobicity, from the proteins with the identified tyrosine residues were lower than those from the whole proteome (Figure 3c). The result suggests that the identified proteins were more hydrophilic, which may facilitate the reaction between tyrosine and the aryl diazonium-alkyne probe in aqueous solution. About 50% of the modified tyrosine residues were annotated as phosphorylation sites based on the information from the PhosphoSitePlus database, consistent with the previous results from sulfur-triazole exchange chemistry (Figure 3d).[27] Next, we investigated the relationship between the modified sites and protein secondary structures. NetsurfP[52] was used to predict the secondary structures of the identified proteins, and the results are shown in Figure S5. Among all the tyrosine residues, the percentage at the coil structure was the highest while the lowest percentage came from the β-strand structure. The disordered structures of coils may increase the chance of the exposure of the tyrosine residues to solvent, and therefore promote the azo-coupling reaction.

Analysis of the quantified tyrosine residues

The distribution of the measured H/L ratios (high concentration versus low concentration) is displayed in Figure 4a. The ratios seem to be higher than those from some previously reported results about tyrosine, lysine and cysteine. The ratio distribution is related to multiple experimental conditions, such as the reaction time and temperature. Furthermore, the higher and lower concentrations play a very important role on the ratio distribution. The quantified tyrosine residues were separated into three groups based on their ratios and then we performed further investigation regarding the differences among these groups.

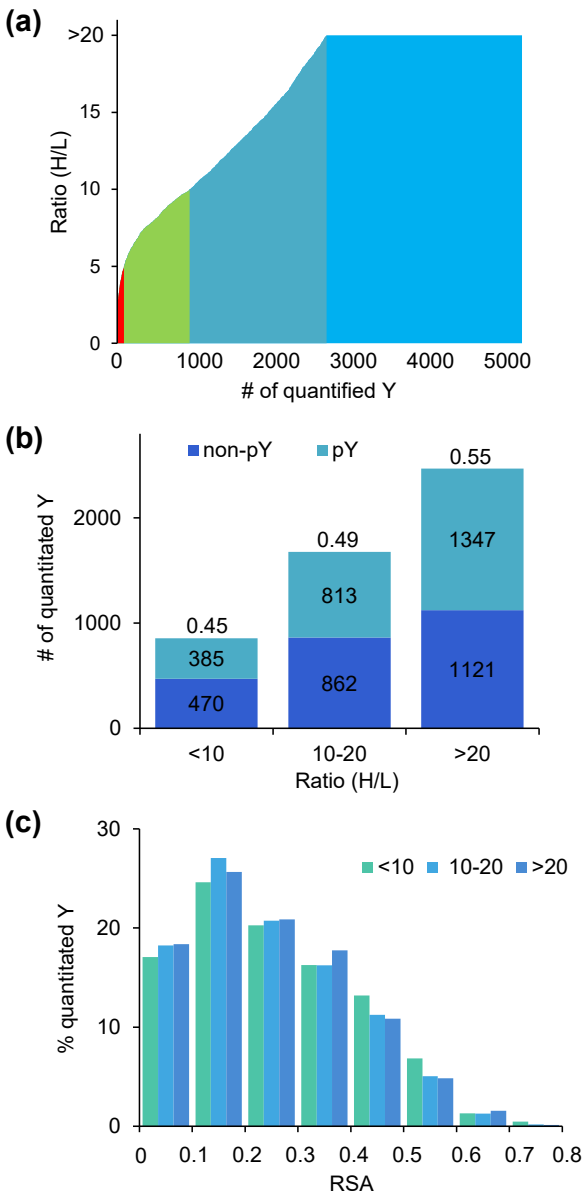

Figure 4. (a) The distribution of the measured ratios (H/L) for the quantified tyrosine residues in MCF7 cells labeled with high- or low-concentration of probe **1**. (b) Comparison of the reported phosphorylation sites (pY) and non-phosphorylation sites (non-pY) among the quantified tyrosine residues in the different ratio ranges. (c) The distribution of the relative surface accessibility (RSA) of the tyrosine sites in the groups with different ratio ranges.

Among all the quantified tyrosine sites, 34% and 49% sites have the ratios of 10-20 and >20, respectively, while only 17% have the ratios of <10, of which 75 sites (about 9%) are <5 (Figure 4a, Table S3). The H/L ratio is correlated with the reactivity of the tyrosine residues. For example, the site at Y197 of glycogen phosphorylase (PYGB), a critical enzyme in glycogen metabolism, was quantified with a ratio of 5.9 (Figure S6). Previous research showed that Y197 forms a hydrogen bond with AMP (adenosine monophosphate), which promotes the binding of AMP to PYGB and then activates its catalytic function.[53] Another two tyrosine residues on this protein were also quantified, but they have much higher ratios: 26.8 for Y473 and 12.2 for Y821.

The fraction of the quantified tyrosine sites that were reported as phosphorylation sites differed in each category based on the ratios with the smallest fraction from the group with the ratios of <10 (0.45) (Figure 4b). The distribution of the relative surface accessibility (RSA) of the quantified tyrosine sites was similar among the three groups, which indicates that the reactivity of tyrosine is not determined by its accessibility in the aqueous solution, and may be regulated by other factors such as the microenvironment created by nearby amino acids. For the tyrosine residues in the three groups, the acidic amino acids (D and E) were enriched near the quantified tyrosine sites (± 1 and ± 2 positions) (Figures S7a-c). However, as the ratio increased, the basic amino acids (K or R) at the positions of +5 and +6 became overrepresented. We reason that this unique pattern that creates specific microenvironment may lead to different reactivities of tyrosine on proteins. The carboxyl group can form a hydrogen bond with tyrosine, which may increase the reactivity of tyrosine. However, when a basic amino acid residue is around, the electrostatic effect between protonated K or R and the carboxyl group becomes dominant, which may prevent the formation of the hydrogen bond between the carboxyl group and tyrosine (Figure S7d). Therefore, it may result in the lower reactivity of the tyrosine.

Functional analysis of proteins with the quantified tyrosine residues

Reactive amino acid residues are central to the biological functions of proteins because they are essential for many important activities of proteins such as catalytic activity and ligand binding. Here, we identified >700 proteins containing the quantified tyrosine sites with an H/L ratio of <10 (Table S3c). Functional annotation revealed that these proteins are highly enriched in different activities including nucleotide binding, enzyme binding, drug binding, protein kinase binding, and hydrolase activity (Figure 5a). Particularly, about 30% of the proteins have catalytic activity.

To further analyze the protein functions, we performed network clustering using clusterMaker in Cytoscape, which analyzes the protein complexes in a network. Protein complexes formed through protein-protein interactions (PPIs) are very important to a cellular system. The MCL clustering algorithm generated over 50 protein complexes (Figure S8a). The largest cluster contains proteins related to mRNA splicing. The generated clusters are involved in different pathways including aminoacyl-tRNA biosynthesis, endocytosis, insulin signaling pathway. We also compared the generated protein complexes to CORUM, a database deposited with manually curated protein complexes, to further evaluate the biological functions of the network. Two examples are shown in Figure S8b, and the proteins containing the tyrosine sites with a ratio of <10 were reported to participate in EIF3 complex and proteasome.

For the tyrosine residues with a lower H/L ratio (<5), their reactivity is supposed to be higher. Proteins with these tyrosine residues are correlated with different types of activities and functions (Figure 5b), with the dominant groups having catalytic activity (45%) and binding (34%). They were also found to participate in different types of protein complexes (Figure S8b).

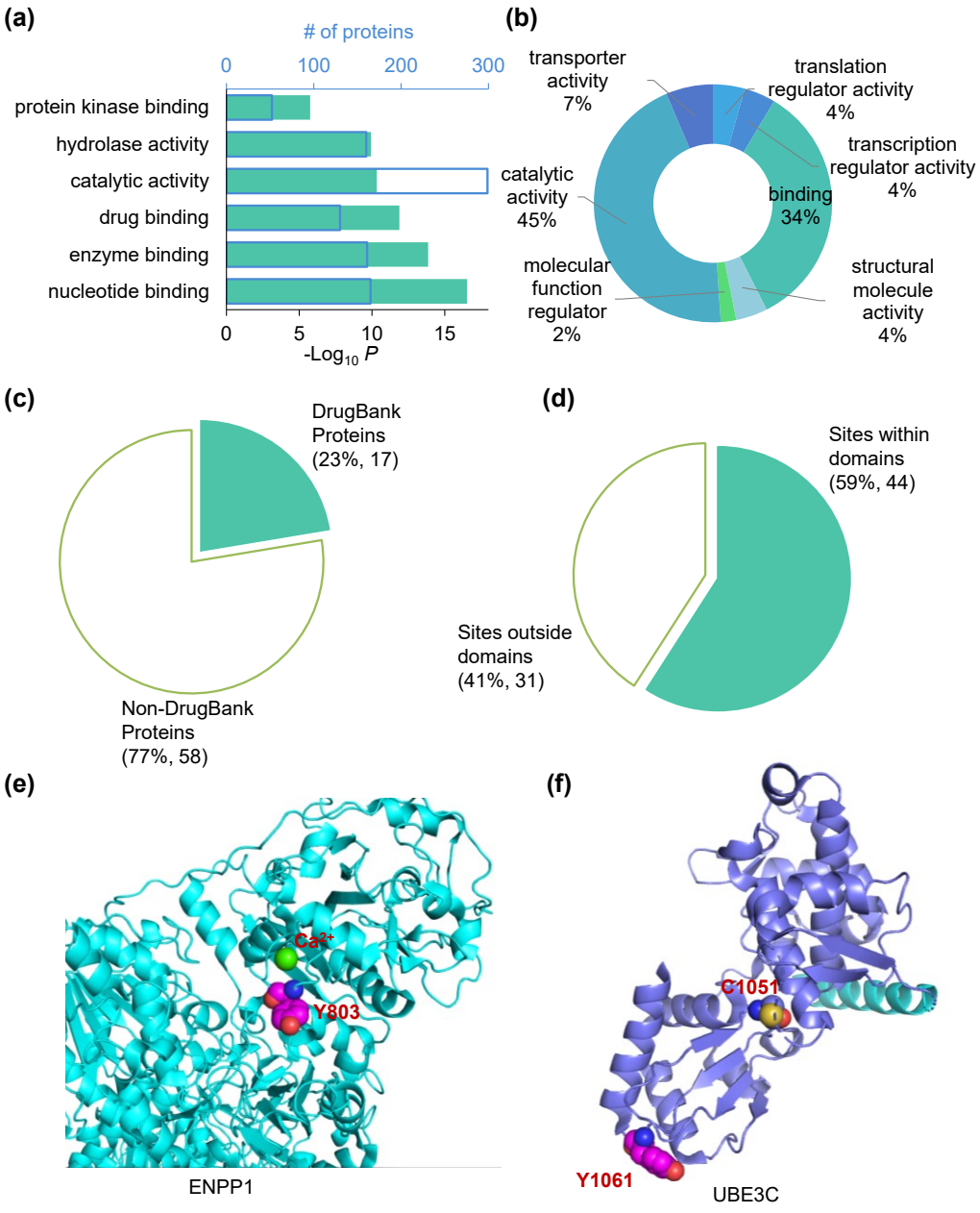

Figure 5. (a) Clustering based on molecular function for the proteins containing the quantified tyrosine residues with a ratio of <10. (b) Functional annotation of the proteins containing the quantified sites with a ratio of <5. (c) The fraction of proteins possessing unique tyrosine residues with a ratio of <5 found in DrugBank. (d) Domain analysis of the tyrosine sites with a ratio of <5. (e) and (f) Example structures of the proteins ENPP1 (PDB: 4B56) and UBE3C (PDB: 6K2C) with reactive tyrosine residues (H/L<5).

For proteins harboring the tyrosine residues with a ratio of <5, 23% were found in DrugBank, half of which are enzymes (Figure 5c). Non-DrugBank proteins (77%) came from various protein classes including transporters, transcription factors, and cytoskeletal proteins, which are usually difficult to be targeted by small molecules. Of these sites with the ratio of <5, 59% were located in different protein domains (Figure 5d), including binding, enzyme, and protein-protein interaction domains, which indicated that these sites may participate in the mediation of protein functions. We also examined how the reactive sites may affect protein functions. Ectonucleotide pyrophosphatase/phosphodiesterase-1 (ENPP1) is a calcium- and zinc-dependent enzyme that can hydrolyze phosphodiester or pyrophosphate bonds.[54] Y803 (Figure 5e) was located in nuclease-like domain of ENPP1 and found to be within the binding pocket for Ca^{2+} (<5Å).[55] A reactive site (Y1061, Figure 5f) was identified on the E3 ligase catalytic domain of Ubiquitin-protein ligase E3C (UBE3C), an enzyme involved in protein ubiquitination. Although the site is not near the catalytic site of C1051 (>5Å), it may affect the UBE3C's function through allosteric regulation since both Y1061 and C1051 are located on the same domain.[56] Overall, the tyrosine residues with high reactivity were found on a wide range of protein classes, which provides valuable information for future development of covalent drugs to manipulate protein activities and expands the scope of drug targets.

CONCLUSIONS

The tyrosine residue is involved in diverse protein functions including ligand binding, catalysis, and cell signaling. The functions of tyrosine are often correlated with its intrinsic reactivity. A comprehensive investigation of the reactivity of tyrosine in the whole proteome will facilitate the

development of chemical probes and covalent drugs to selectively tune the biological activities of proteins. In this work, we developed a new method integrating azo coupling, bioorthogonal chemistry, and multiplexed proteomics to globally study the tyrosine residues in the human proteome. Based on the azo-coupling reaction between aryl diazonium salt and the tyrosine residue, the probe can specifically target the tyrosine residues. After the reaction, tagged tyrosine containing peptides were selectively enriched using bioorthogonal chemistry, and a small tag on the peptides from the cleavage perfectly fits for site-specific analysis by MS. Over 5,000 tyrosine residues were quantified in MCF7 cells in the biological triplicate experiments. Although most of them were found to display a high H/L value (>10) in the concentration-dependent experiment, the quantified tyrosine residues with a low H/L ratio (<5) were found to be located in proteins with different functions including catalytic activity, kinase binding, and hydrolase activity. In combination with multiplexed proteomics, this method enables the global profiling of tyrosine in the proteome, resulting in a better understanding of protein functions and the development of covalent drugs to regulate protein activity.

ASSOCIATED CONTENT

Supporting information

The supporting information including Supplementary Figures S1-S8 and Supplementary Tables S1-S3 is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

E-mail: ronghu.wu@chemistry.gatech.edu

ACKNOWLEDGMENT

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM127711.

References

(1) Aebersold, R.; Mann, M., *Nature* **2016**, *537*(7620): 347-355.
(2) Cravatt, B. F.; Simon, G. M.; Yates, J. R., *Nature* **2007**, *450*(7172): 991-1000.
(3) Altelaar, A. F. M.; Munoz, J.; Heck, A. J. R., *Nat. Rev. Genet.* **2013**, *14*(1): 35-48.
(4) Chen, W.; Smeekens, J. M.; Wu, R., *Chem. Sci.* **2016**, *7*(2): 1393-1400.
(5) Arul, A. B.; Robinson, R. A. S., *Anal. Chem.* **2019**, *91*(1): 178-189.
(6) Huang, M.; Wang, Y. S., *Mass Spectrom. Rev.* **2021**, *40*(3): 215-235.
(7) Woo, C. M.; Felix, A.; Byrd, W. E.; Zuegel, D. K.; Ishihara, M.; Azadi, P.; Iavarone, A. T.; Pitteri, S. J.; Bertozzi, C. R., *J. Proteome Res.* **2017**, *16*(4): 1706-1718.
(8) Suttapitugsakul, S.; Sun, F.; Wu, R., *Anal. Chem.* **2020**, *92*(1): 267-291.
(9) Yang, Y.; Franc, V.; Heck, A. J. R., *Trends Biotechnol.* **2017**, *35*(7): 598-609.
(10) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M., *Cell* **2006**, *127*(3): 635-648.
(11) Wu, R.; Haas, W.; Dephoure, N.; Huttlin, E. L.; Zhai, B.; Sowa, M. E.; Gygi, S. P., *Nat. Methods* **2011**, *8*(8): 677-683.
(12) Aguilar, H. A.; Iliuk, A. B.; Chen, I. H.; Tao, W. A., *Nat. Protoc.* **2020**, *15*(1): 161-180.
(13) Taylor, B. C.; Young, N. L., *Biochem. J.* **2021**, *478*(3): 511-532.
(14) Udeshi, N. D.; Svinkina, T.; Mertins, P.; Kuhn, E.; Mani, D. R.; Qiao, J. W.; Carr, S. A., *Mol. Cell. Proteomics* **2013**, *12*(3): 825-831.
(15) Li, Y. N.; Evers, J.; Luo, A.; Erber, L.; Postler, Z.; Chen, Y., *Angew. Chem.-Int. Edit.* **2019**, *58*(2): 537-541.
(16) Weerapana, E.; Wang, C.; Simon, G. M.; Richter, F.; Khare, S.; Dillon, M. B. D.; Bachovchin, D. A.; Mowen, K.; Baker, D.; Cravatt, B. F., *Nature* **2010**, *468*: 790.
(17) Fomenko, D. E.; Xing, W.; Adair, B. M.; Thomas, D. J.; Gladyshev, V. N., *Science* **2007**, *315*(5810): 387.
(18) Wang, J.; Liu, Y.; Liu, Y.; Zheng, S.; Wang, X.; Zhao, J.; Yang, F.; Zhang, G.; Wang, C.; Chen, P. R., *Nature* **2019**.
(19) Hacker, S. M.; Backus, K. M.; Lazear, M. R.; Forli, S.; Correia, B. E.; Cravatt, B. F., *Nat. Chem.* **2017**, *9*: 1181.
(20) Ma, N.; Hu, J.; Zhang, Z.-M.; Liu, W.; Huang, M.; Fan, Y.; Yin, X.; Wang, J.; Ding, K.; Ye, W.; Li, Z., *J. Am. Chem. Soc.* **2020**, *142*(13): 6051-6059.
(21) Bach, K.; Beerkens, B. L. H.; Zanon, P. R. A.; Hacker, S. M., *ACS Cent. Sci.* **2020**, *6*(4): 546-554.
(22) Lin, S.; Yang, X.; Jia, S.; Weeks, A. M.; Hornsby, M.; Lee, P. S.; Nichiporuk, R. V.; Iavarone, A. T.; Wells, J. A.; Toste, F. D.; Chang, C. J., *Science* **2017**, *355*(6325): 597.
(23) Mukherjee, H.; Grimster, N. P., *Curr. Opin. Chem. Biol.* **2018**, *44*: 30-38.
(24) Gu, C.; Shannon, D. A.; Colby, T.; Wang, Z.; Shabab, M.; Kumari, S.; Villamor, Joji G.; McLaughlin, Christopher J.; Weerapana, E.; Kaiser, M.; Cravatt, Benjamin F.; van der Hoorn, Renier A. L., *Chem. & Biol.* **2013**, *20*(4): 541-548.
(25) Koide, S.; Sidhu, S. S., *ACS Chem. Biol.* **2009**, *4*(5): 325-334.
(26) Mortenson, D. E.; Brighty, G. J.; Plate, L.; Bare, G.; Chen, W.; Li, S.; Wang, H.; Cravatt, B. F.; Forli, S.; Powers, E. T.; Sharpless, K. B.; Wilson, I. A.; Kelly, J. W., *J. Am. Chem. Soc.* **2018**, *140*(1): 200-210.
(27) Hahm, H. S.; Toroitich, E. K.; Borne, A. L.; Brulet, J. W.; Libby, A. H.; Yuan, K.; Ware, T. B.; McCloud, R. L.; Ciancone, A. M.; Hsu, K.-L., *Nat. Chem. Biol.* **2020**, *16*(2): 150-159.
(28) Brulet, J. W.; Borne, A. L.; Yuan, K.; Libby, A. H.; Hsu, K.-L., *J. Am. Chem. Soc.* **2020**.
(29) Sengupta, S.; Chandrasekaran, S., *Org. Biomol. Chem.* **2019**, *17*(36): 8308-8329.
(30) Addy, P. S.; Erickson, S. B.; Italia, J. S.; Chatterjee, A., *J. Am. Chem. Soc.* **2017**, *139*(34): 11670-11673.

(31) Gavrilyuk, J.; Ban, H.; Nagano, M.; Hakamata, W.; Barbas, C. F., *Bioconjug. Chem.* **2012**, *23*(12): 2321-2328.
(32) Hooker, J. M.; Kovacs, E. W.; Francis, M. B., *J. Am. Chem. Soc.* **2004**, *126*(12): 3718-3719.
(33) Schlick, T. L.; Ding, Z.; Kovacs, E. W.; Francis, M. B., *J. Am. Chem. Soc.* **2005**, *127*(11): 3718-3723.
(34) Mo, F.; Dong, G.; Zhang, Y.; Wang, J., *Org. Biomol. Chem.* **2013**, *11*(10): 1582-1593.
(35) Evrard, D.; Lambert, F.; Policar, C.; Balland, V.; Limoges, B., *Chem. Eur. J.* **2008**, *14*(30): 9286-9291.
(36) Ma, X.; Herzon, S. B., *Beilstein J. Org. Chem.* **2018**, *14*: 2259-2265.
(37) Xiao, H.; Wu, R., *Anal. Chem.* **2017**, *89*(6): 3656-3663.
(38) Eng, J. K.; McCormack, A. L.; Yates, J. R., *J. Am. Soc. Mass Spectrom.* **1994**, *5*(11): 976-989.
(39) Elias, J. E.; Gygi, S. P., *Nat. Methods* **2007**, *4*: 207.
(40) Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J., *Nat. Methods* **2007**, *4*(11): 923-925.
(41) Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P., *Nat. Biotechnol.* **2006**, *24*(10): 1285-1292.
(42) Hornbeck, P. V.; Zhang, B.; Murray, B.; Kornhauser, J. M.; Latham, V.; Skrzypek, E., *Nucleic Acids Res.* **2014**, *43*(D1): D512-D520.
(43) O'Shea, J. P.; Chou, M. F.; Quader, S. A.; Ryan, J. K.; Church, G. M.; Schwartz, D., *Nat. Methods* **2013**, *10*: 1211.
(44) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T., *Genome Res.* **2003**, *13*(11): 2498-2504.
(45) Szklarczyk, D.; Morris, J. H.; Cook, H.; Kuhn, M.; Wyder, S.; Simonovic, M.; Santos, A.; Doncheva, N. T.; Roth, A.; Bork, P.; Jensen, L. J.; von Mering, C., *Nucleic Acids Res.* **2016**, *45*(D1): D362-D368.
(46) Mi, H.; Muruganujan, A.; Ebert, D.; Huang, X.; Thomas, P. D., *Nucleic Acids Res.* **2018**, *47*(D1): D419-D426.
(47) Morris, J. H.; Apeltsin, L.; Newman, A. M.; Baumbach, J.; Wittkop, T.; Su, G.; Bader, G. D.; Ferrin, T. E., *BMC Bioinformatics* **2011**, *12*(1): 436.
(48) Giurgiu, M.; Reinhard, J.; Brauner, B.; Dunger-Kaltenbach, I.; Fobo, G.; Frishman, G.; Montrone, C.; Ruepp, A., *Nucleic Acids Res.* **2018**, *47*(D1): D559-D563.
(49) Pandurangan, A. P.; Stahlhacke, J.; Oates, M. E.; Smithers, B.; Gough, J., *Nucleic Acids Res.* **2018**, *47*(D1): D490-D494.
(50) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., *Nucleic Acids Res.* **2000**, *28*(1): 235-242.
(51) Patrick R. A., Z.; Fengchao, Y.; Patricia, M.; Lisa, L.; Michael, Z.; Kristina, K.; Dario, M.; Patrick, R.; Thomas E., M.; Marko, C.; Christopher, C.; Kathrin, L.; F. Dean, T.; Alexey I., N.; Stephan M., H., Profiling the proteome-wide selectivity of diverse electrophiles. 2021.
(52) Petersen, B.; Petersen, T. N.; Andersen, P.; Nielsen, M.; Lundegaard, C., *BMC Struct. Biol.* **2009**, *9*(1): 51.
(53) Mathieu, C.; de la Sierra-Gallay, I. L.; Duval, R.; Xu, X.; Cocaign, A.; Léger, T.; Woffendin, G.; Camadro, J.-M.; Etchebest, C.; Haouz, A.; Dupret, J.-M.; Rodrigues-Lima, F., *J. Biol. Chem.* **2016**, *291*(35): 18072-18083.
(54) Onyedibe, K. I.; Wang, M.; Sintim, H. O., *Molecules* **2019**, *24*(22): 4192.
(55) Jansen, S.; Perrakis, A.; Ulens, C.; Winkler, C.; Andries, M.; Joosten, Robbie P.; Van Acker, M.; Luyten, Frank P.; Moolenaar, Wouter H.; Bollen, M., *Structure* **2012**, *20*(11): 1948-1959.
(56) Singh, S.; Sivaraman, J., *Biochem. J.* **2020**, *477*(5): 905-923.

Table of Contents