OXFORD

Briefings in Bioinformatics, 00(00), 2020, 1–12

doi: 10.1093/bib/bbaa038
Method Review

GPS-Palm: a deep learning-based graphic presentation system for the prediction of S-palmitoylation sites in proteins

Wanshan Ning†, Peiran Jiang†, Yaping Guo, Chenwei Wang, Xiaodan Tan, Weizhi Zhang, Di Peng and Yu Xue

Corresponding author: Yu Xue, Key Laboratory of Molecular Biophysics of Ministry of Education, Hubei Bioinformatics and Molecular Imaging Key Laboratory, Center for Artificial Intelligence Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China; Huazhong University of Science and Technology Ezhou Industrial Technology Research Institute, Ezhou 436044, China. Tel: +86-27-87793903; Fax: +86-27-87793172; E-mail: xueyu@hust.edu.cn
†These authors contributed equally to this work.

Abstract

As an important reversible lipid modification, S-palmitoylation mainly occurs at specific cysteine residues in proteins, participates in regulating various biological processes and is associated with human diseases. Besides experimental assays, computational prediction of S-palmitoylation sites can efficiently generate helpful candidates for further experimental consideration. Here, we reviewed the current progress in the development of S-palmitoylation site predictors, as well as training data sets, informative features and algorithms used in these tools. Then, we compiled a benchmark data set containing 3098 known S-palmitoylation sites identified from small- or large-scale experiments, and developed a new method named data quality discrimination (DQD) to distinguish data quality weights (DQWs) between the two types of the sites. Besides DQD and our previous methods, we encoded sequence similarity values into images, constructed a deep learning framework of convolutional neural networks (CNNs) and developed a novel algorithm of graphic presentation system (GPS) 6.0. We further integrated nine additional types of sequence-based and structural features, implemented

Wanshan Ning is a PhD student at Huazhong University of Science and Technology. His major research interest is focused on using artificial intelligence methods to analyze sequence, omics and imaging data. He developed a new hybrid-learning architecture named HybridSucc for predicting general and species-specific succinylation sites. He was a major developer of DrLLPS, a comprehensive database containing curated proteins associated with liquid-liquid phase separation. He also constructed a tool named WocEA for the visualization of enrichment analyses in word clouds.
Peiran Jiang is an undergraduate student at Huazhong University of Science and Technology. He mainly focuses on the prediction of S-palmitoylation sites by using deep learning algorithms.
Yaping Guo is a PhD student at Huazhong University of Science and Technology. She was a major developer of two integrative databases, including DrLLPS and iEKPD 2.0 for protein kinases, protein phosphatases and proteins containing phosphoprotein-binding domains in eukaryotes. Her major research interest is the development of PTM-related data resources.
Chenwei Wang is a PhD student at Huazhong University of Science and Technology. He mainly focuses on the development of new algorithms to predict functional PTM events from multi-omic data.
Xiaodan Tan is a master student at Huazhong University of Science and Technology. She mainly focuses on the development of high-throughput methods to identify functional phosphorylation events, by using deep learning algorithms.
Weizhi Zhang is a PhD student at Huazhong University of Science and Technology. He is working on the prediction of proteins containing important motifs involved in autophagy.
Di Peng is a postdoc scientist at Huazhong University of Science and Technology. His major research interests are focused on the experimental discovery of new PTM regulators, substrates and sites, based on computational predictions. He was a major developer of iEKPD 2.0.
Yu Xue is a professor at Huazhong University of Science and Technology. He has started to work in the field of PTM Bioinformatics since 2004. He is interested in using both artificial intelligence and experimental approaches to exploit how functional PTM events can be precisely orchestrated to regulate various biological processes, such as autophagy, circadian and cell fate determination.
Submitted: 11 December 2019; **Received (in revised form):** 19 February 2020

parallel CNNs (pCNNs) and designed a new predictor called GPS-Palm. Compared with other existing tools, GPS-Palm showed a >31.3% improvement of the area under the curve (AUC) value (0.855 versus 0.651) for general prediction of S-palmitoylation sites. We also produced two species-specific predictors, with corresponding AUC values of 0.900 and 0.897 for predicting human- and mouse-specific sites, respectively. GPS-Palm is free for academic research at http://gpspalm.biocuckoo.cn/.

Key words: S-palmitoylation; post-translational modification; data quality discrimination; convolutional neural networks; graphic presentation system; parallel CNNs

Introduction

As an important and special class of post-translational modifications (PTMs), lipid modifications mainly comprise S-palmitoylation (C16), N-myristoylation (C14), S-farnesylation (C15), S-geranylgeranylation (C20), cholesterylation and glycosylphosphatidylinositol (GPI)-anchor, depending on the type of lipids covalently attached to modified substrate proteins [1, 2]. Unlike other tethering lipid modifications, S-palmitoylation reversibly adds one or multiple palmitoyl moieties to internal cysteine residues in proteins through the thioesterification reaction [3–7]. S-palmitoylation effectively increases the hydrophobicity of protein surfaces to dynamically regulate membrane–protein interactions [1, 8] and participates in regulating a broad spectrum of biological processes, such as signal transduction [2, 7], neuronal transmission [3], metabolism [9], autophagy [10] and immunological response [11]. In addition, dysregulation of S-palmitoylation is associated with numerous human diseases such as cancer [11, 12], neurodegenerative disorders [13] and diabetes [14]. Although the biological importance of protein S-palmitoylation has been gradually recognized in recent years, its underlying mechanisms are still unclear.

Identification of palmitoylated substrates with exact sites is fundamental for understanding the molecular mechanisms and regulatory roles of S-palmitoylation. Conventionally, S-palmitoylated proteins were identified by metabolically labeling with [^{3}H] palmitate *in vivo* [3, 15]. Owing to the lack of clear sequence motifs for S-palmitoylation recognition, pinpointing exact sites in substrates was labor-intensive and tedious [3, 4, 7, 15]. Advances in mass spectrometry (MS)-based proteomics technology have enabled the detection of an amount of palmitoylated proteins and sites [3, 9, 16–19]. In 2006, a large-scale profiling identified 47 S-palmitoylated proteins in *Saccharomyces cerevisiae*, by coupling the acyl-biotin exchange (ABE) method to MS [3]. Later, a new assay of resin-assisted capture (RAC) was established to purify palmitoylated proteins and increase the sensitivity for identifying palmitoylated peptides [9, 18]. It should be noted that when a palmitoylated peptide contains multiple cysteine residues, it would be difficult to clearly determine the S-palmitoylation sites [9]. Thus, computational predictions of S-palmitoylation sites with bioinformatic approaches can efficiently tackle this problem and generate useful candidates for further experiments.

Here, we reviewed the mainstream computational methods and tools for the prediction of S-palmitoylation sites, including CSS-Palm 1.0 [20], NBA-Palm 1.0 [21], CSS-Palm 2.0 [22], CKSAAP-Palm [23], PPWMs [24], IFS-Palm [25], WAP-Palm [26], PalmPred [27], SeqPalm [28], GPS-Lipid [29] and MDD-Palm [30] (Supplementary Table S1). Through the literature biocuration and public database integration, we compiled a large benchmark data set containing 3098 unique and nonhomologous S-palmitoylation sites in 1618 proteins, which were experimentally identified from small- or large-scale studies (Figure 1A, Supplementary Table S2). Then, we developed a new method named data quality discrimination (DQD) to measure data quality weights (DQWs), and observed that small-scale sites had significantly higher DQWs than large-scale sites. We incorporated DQD into our recently developed group-based prediction system (GPS) 5.0 algorithm, which implemented two additional methods of position weight determination (PWD) and scoring matrix optimization (SMO) for performance improvement (Figure 1B) and achieved an area under the curve (AUC) value of 0.749 for predicting S-palmitoylation sites.

Inspired by DeepVariant, a pioneering tool that encoded genomic sequencing data into images for calling genetic variants [31], we further designed a new strategy of number-to-image transformation (NIT) to transform numerical sequence similarity values into images (Figure 1C), which were then inputted into a deep learning framework of 11-layer convolutional neural networks (CNNs) for model training. Together with all improvements, we renamed this update of GPS algorithm as graphic presentation system 6.0, with an increased AUC value of 0.806. Additionally, we used NIT to encode six additional types of sequence-derived features including pseudo amino acid composition (PseAAC), composition of k-spaced amino acid pairs (CKSAAP), orthogonal binary coding (OBC), physicochemical properties in the Amino Acid index database (AAindex), autocorrelation functions (ACF) and position-specific scoring matrix (PSSM), and three types of structural features including accessible surface area (ASA), secondary structure (SS) and backbone torsion angles (BTA) [20–30, 32] (Figure 1C, Supplementary Methods, Supplementary Table S3). Parallel CNNs (pCNNs) were implemented for training and for integrating up to 2835 individual features (Figure 1D), and then we developed a new tool called GPS-Palm. Through a comparison with other existing tools, GPS-Palm exhibited a >31.34% improvement of AUC value (0.855 versus 0.651) for general prediction of S-palmitoylation sites. In addition, GPS-Palm also generated two species-specific models for predicting human- and mouse-specific sites, with AUC values of 0.900 and 0.897, respectively. Taken together, we anticipate that GPS-Palm might be a helpful tool to analyze S-palmitoylation, and all approaches used in this study can be extended to predict other types of PTM sites. The local packages of GPS-Palm were implemented in Python and can be downloaded at: http://gpspalm.biocuckoo.cn/download.php.

Methods

Data collection and preparation.

First, we searched PubMed with a number of keywords, such as 'protein palmitoylation,' 'palmitoylation,' 'cysteine palmitoylation,' 'S-palmitoylation' and 'palmitoylated.' The full texts of all retrieved papers were carefully checked, and we manually

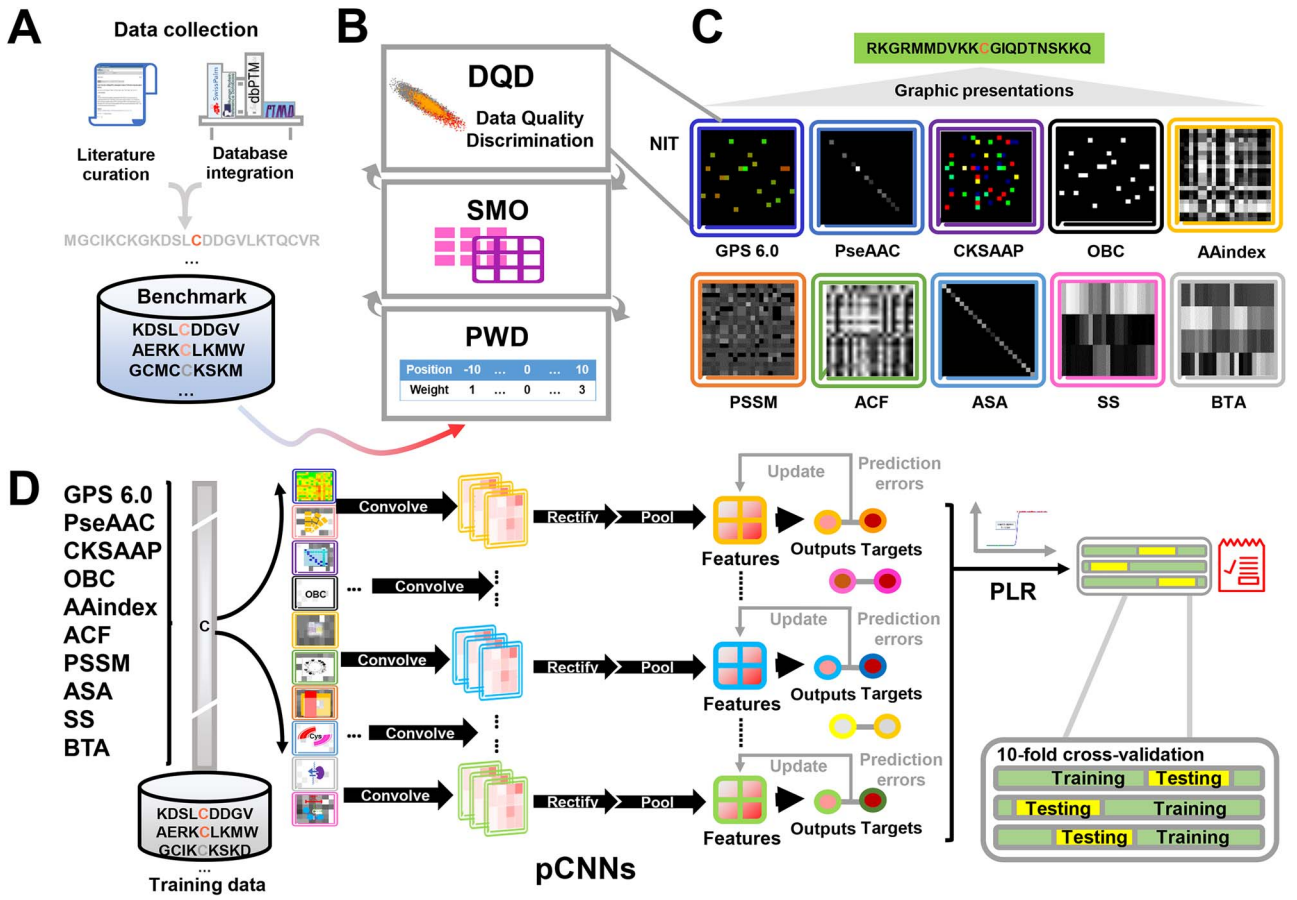

Figure 1. The experimental procedure of the study. (A) From the scientific literature and four public databases including dbPTM [33], SwissPalm [34], HPRD [35] and PTMD [36], we collected 5978 unique S-palmitoylation sites in 2840 proteins, after redundancy clearance. After homology elimination, the final benchmark data set contained 3098 known sites in 1618 substrates. (B) We developed the GPS 6.0 by implementing two additional methods of DQD and NIT-based CNNs, which were combined with PWD and SMO in the GPS 5.0 algorithm to improve the prediction performance. The basic scoring strategy was reserved. (C) Besides GPS, we used NIT to encode six additional features, including six sequence-based features of PseAAC, CKSAAP, OBC, AAindex, ACF and PSSM, and three structural features of ASA, SS and BTA [20–30, 32]. (D) To develop GPS-Palm, we implemented a deep learning framework of pCNNs to integrate 2835 individual features for training a single model. The 10-fold cross-validations were performed to evaluate the accuracy.

collected 5849 experimentally identified S-palmitoylation sites through the literature biocuration. To avoid missing any data, we further obtained 5183 known S-palmitoylation sites from four public databases, including dbPTM [33], SwissPalm [34], HPRD [35] and PTMD [36]. We merged the two data sets together, and mapped S-palmitoylation sites to primary protein sequences downloaded from the UniProt database [37] to pinpoint the exact modification positions. In total, we obtained 5978 nonredundant S-palmitoylation sites in 2840 proteins.

Before training, homologous sites should be eliminated to avoid overfitting. Thus, we used CD-HIT [38], a program for clustering similar biological sequences, to cluster palmitoylated protein sequences, with a threshold of 40% sequence similarity. If two proteins are palmitoylated at the same positions with a >40% sequence identity, only one representative sequence was retained. Then, we defined a palmitoylation site peptide PSP(m, n) as a cysteine residue flanked by m residues upstream and n residues downstream. Because too many parameters would be determined and optimized, here we chose PSP(10, 10) for a rapid training. As previously described [22], PSP(10, 10) items derived from known S-palmitoylation sites were taken as positive data, whereas PSP(10, 10) peptides around nonpalmitoylated cysteine residues in the same proteins were regarded as negative data. Finally, we constructed a high-quality benchmark data set, containing 3098 positive sites and 18 992 negative sites from 1618 substrates (Supplementary Table S2). From the data set, we found that known S-palmitoylation sites were experimentally characterized by different approaches. Thus, we simply took known sites verified by conventionally biochemical assays as 'small-scale sites,' whereas the remaining sites only detected by high-throughput MS were taken as 'large-scale sites.' For each known palmitoylated protein, its UniProt accession number, protein sequence, S-palmitoylation sites, experimental type, organisms and PubMed IDs (PMIDs) of original references were provided and could be accessed at http://gpspalm.biocuckoo.cn/userguide.php.

Performance evaluation

To evaluate the prediction accuracy of various computational methods, we calculated five measurements, including sensitivity (Sn), specificity (Sp), accuracy (Ac), precision (Pr) and Mathew correlation coefficient (MCC).

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \mathrm{Ac} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}, \mathrm{Pr} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FN}\times\mathrm{FP})}{\sqrt{(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}} \quad (1)$$

For each method, the 10-fold cross-validation was repeatedly performed 20 times, and the average Sn, Sp, Ac, Pr and MCC values were calculated. Then, the receiver operating characteristic (ROC) curve was illustrated based on final Sn and 1-Sp scores, and the average AUC value was computed.

The GPS 6.0 algorithm

In 2004, we developed the GPS 1.0 algorithm with the full name of group-based phosphorylation site predicting and scoring platform to measure the local sequence similarity between a given phosphorylation site (p-site) peptide and all known p-sites in positive data, based on a hypothesis of similar peptides potentially sharing similar properties [39]. Later, we renamed it into group-based prediction system [40], whereas the basic scoring strategy was never changed in all versions of GPS algorithms, and the latest GPS 5.0 algorithm implemented two additional approaches including PWD and SMO for performance improvement (http://gps.biocuckoo.cn/).

In the scoring strategy, the similarity score between two PSP(10, 10) peptides A and B was formulated as below:

$$S(A,B)=\sum_{1\le i\le 20} W_i M(A[i],B[i]) \tag{2}$$

Here, $W_i$was the weight of position i, and the $M(A[i], B[i])$ was the substitution score of the amino acid pair $(A[i], B[i])$ aligned at position i. The substitution score is symmetrical with $M(a, b) = M(b, a)$. Initially, all position values in W were taken as 1, while the BLOSUM62 matrix was used as the starting matrix. Then in the step of performance improvement, PWD and SMO were iteratively adopted to optimize the trainable parameters in W and M, respectively, until the average AUC value of the 10-fold cross-validations was not increased any longer. In GPS 5.0, the original penalized logistic regression (PLR) algorithm with the ridge (L2) penalty was used for training models.

In this study, we further developed a new method named DQD, and defined the average similarity score PS(A) between a given PSP(10, 10) A and the whole positive data set P with T_+ peptides as

$$\mathrm{PS(A)}=\frac{1}{T_+}\sum_{j=1}^{T_+} S(A,P_j)\,p\mathrm{DQW}_j \tag{3}$$

where positive DQW_j ($p\mathrm{DQW}_j$) was the DQW value of P_j in P, and pDQW was a weight vector of trainable DQWs for the positive data set. Also, we defined the average similarity score NS(A) between a given PSP(10, 10) A and the whole negative data set N with T_- peptides as

$$\mathrm{NS(A)}=\frac{1}{T_-}\sum_{j=1}^{T_-} S(A,N_j)\,n\mathrm{DQW}_j \tag{4}$$

where negative DQW_j ($n\mathrm{DQW}_j$) was the DQW value of N_j in N, and nDQW was a weight vector of trainable DQWs for the negative data set. All values in pDQW and nDQW were initialized as 1. We developed an improved PLR algorithm (Supplementary Methods), which was used to iteratively optimize all trainable parameters in DQD, PWD and SMO, until the 10-fold cross-validation AUC value was not enhanced any longer.

In order to use CNNs for model training, we designed a new approach named NIT to transform GPS features into images, while both DQD and NIT-based CNNs were incorporated into our previous approaches to develop the GPS 6.0 algorithm. For the positive data set, we first transformed individual PS(A) values into a similarity matrix $\mathrm{Mat}_+(A)$, in which the 21 rows represented 21 types of pseudo amino acids ($A, C, D, \ldots, Y, *$) shown in alphabetical order, and the 20 columns denoted 20 positions in PSP(10, 10) peptide A (from −10 to 10). Central S-palmitoylated cysteine residues were not taken into consideration to avoid overfitting. The matrix was shown as below:

$$\mathrm{Mat}_+(A)=\begin{pmatrix}\mathrm{Mat}_+(A)[A,-10] & \cdots & \mathrm{Mat}_+(A)[A,10]\\ \vdots & \ddots & \vdots\\ \vdots & \ddots & \vdots\\ \mathrm{Mat}_+(A)[*,-10] & \cdots & \mathrm{Mat}_+(A)[*,10]\end{pmatrix}_{21\times 20} \tag{5}$$

In $\mathrm{Mat}_+(A)$, any value $\mathrm{Mat}_+(A)[a, i]$ for an amino acid a in the position i of A could be calculated as below:

$$\mathrm{Mat}_+(A)[a,i]=\frac{1}{T_+(a,i)}\sum_{j=1}^{T_+(a,i)} W_i M(a,A[i])\,p\mathrm{DQW}_j \tag{6}$$

where $T_+(a, i)$ was the number of PSP(10, 10) items in the positive data set with the residue a at position i. Analogously, the similar matrix $\mathrm{Mat}_-(A)$ was also determined between A and the whole negative data set.

For each given PSP(10, 10) peptide A, two similarity matrices $\mathrm{Mat}_+(A)$ and $\mathrm{Mat}_-(A)$ were generated and transformed into an RGB image with two layers. The red and green channels were used for representing $\mathrm{Mat}_+(A)$ and $\mathrm{Mat}_-(A)$, respectively. For the red channel, an element E in the $\mathrm{Mat}_+(A)$ was normalized to 0∼255 as below:

$$E_{\mathrm{Normalized}}=\frac{E-E_{\min}}{E_{\max}-E_{\min}}\times 255 \tag{7}$$

where $E_{\max}$ and $E_{\min}$ were the maximum and minimum elements in the $\mathrm{Mat}_+(A)$, respectively. The same procedure was also conducted to $\mathrm{Mat}_-(A)$ by using the green channel. Two channels were merged to output an intact image, and the blue channel was not utilized. Thus, images of two layers contained both the similarity values of A against the whole positive and negative data sets, respectively. The final graphic presentation of GPS features for a given PSP(10, 10) peptide was an informative 21 × 20-pixel double colored image. The graphic presentations for other types of features were carefully described in Supplementary Methods. The implementation of CNNs and pCNNs for GPS 6.0 and GPS-Palm was described as below.

A deep learning framework of pCNNs

For each feature type, a framework of 11-layer CNNs was adopted, containing one input layer, four pairs of convolutional and pooling layers, one fully connected (dense) layer and one output layer (Figure 2). In the nine hidden layers, neurons were the basic computation units, and both internal feature coding and computational outcome were connected and propagated by neurons inside each layer. The convolutional layers were used for feature extraction and presentation, and a widely used rectified linear unit (ReLU) function was used to activate the outcome of

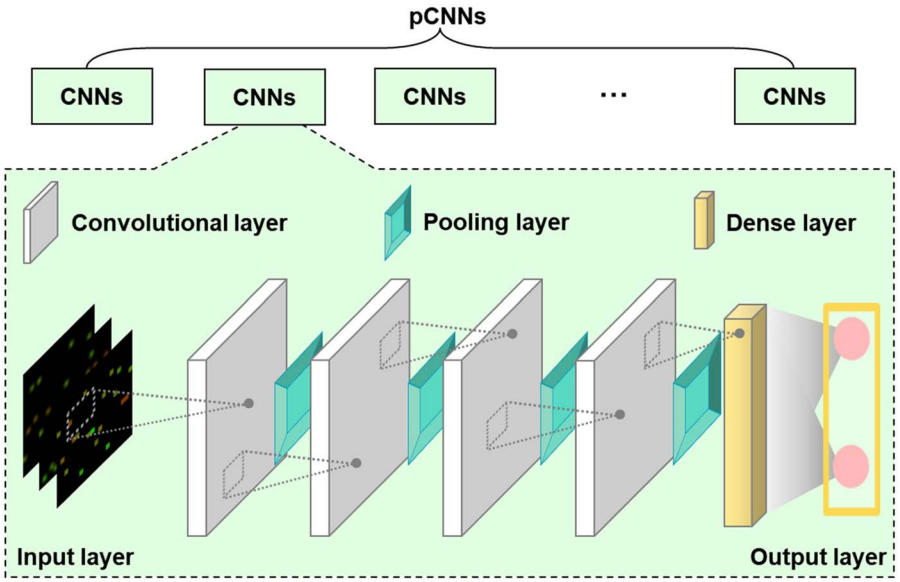

Figure 2. The network architecture of pCNNs. A single CNN model contained 11 layers, including one input layer, four pairs of convolutional and pooling layers for feature presentation and selection, one dense layer for classification and one output layer for generating predictions. ReLU function and max pooling strategy were used in convolutional layers and pooling layers, respectively. Predefined parameters for each framework of pCNNs are shown in Supplementary Table S4.

a neuron and defined as below:

$$\mathrm{ReLU}(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{8}$$

where x was the weighted sum of a neuron.

In the pooling layers, feature selection and information filtering were performed by the max pooling strategy. The last hidden layer was the dense layer for generating prediction outcomes. In order to prevent overfitting that frequently occurs in deep learning algorithms, we used a simple dropout method to randomly select a number of nodes in the dense layer and set their corresponding scores to 0 if the *Ac* value went up. In the output layer, two sigmoid nodes were set to finally calculate the score for a given PSP(10, 10) peptide shown as below:

$$\mathrm{Score}(y) = \frac{1}{1 + e^{-y}} \tag{9}$$

where y was the input of the sigmoid node derived from the dense layer and Score(y) was a 0–1 value to represent the probability of a PSP(10, 10) to be a real *S*-palmitoylation site from a single CNN model.

To integrate the 10 CNNs, the input layer first received the 10 images, in which the 10 different types of features were graphically represented for each PSP(10, 10) peptide. Each image representing one feature type entered a CNN model, and the 10 CNN models comprised the pCNNs. Then, the update of parameters in convolutional, pooling and dense layers was continuously performed, until the errors between outputs and targets were not decreased any longer.

Finally, 10 Score(y) values generated from the 10 CNNs were denoted as the secondary vector, and integrated by the improved PLR algorithm. The final P_{score} for the given PSP(10, 10) peptide was calculated as below:

$$P_{score} = \sum_{i=1}^{10} \mathrm{Score}(y)_i \times w_i \tag{10}$$

where P_{score} was a 0–1 value to denote the final probability of a PSP(10, 10) to be a real *S*-palmitoylation site, and w_i was the weight of Score(y) derived from the ith graphic presentation. The final pCNN models were determined based on the highest AUC value of the 10-fold cross-validations, by using the benchmark data set.

For model training, we used a lab computer with an Intel(R) Core™ i7-6700K@ 4.00 GHz central processing unit (CPU), 32 GB of RAM and a NVIDIA GeForce GTX 960 core. The Keras version 2.0.4 (http://github.com/fchollet/keras), a highly useful neural networks API that was written in Python and developed based on the tensorflow 1.2.0, was adopted for a rapid parallel computing. The Adam optimizer in Keras was adopted, by using parameters of 0.001 for learning rate, 0.99 for the first exponential decay rate, 0.999 for the second exponential decay rate and 256 for mini-batch size. In each framework of pCNNs for predicting general or species-specific sites, predefined parameters including sizes of the 11 layers, dropout ratio and number of iterations (epochs) were shown (Supplementary Table S4).

Development of the GPS-Palm software packages

GPS-Palm was written in Python 3.6 with PyQt 5.0 (https://sourceforge.net/projects/pyqt/). For convenience, local packages were constructed by Advanced Installer (Professional License, https://www.advancedinstaller.com/) to support three major operating systems, including Windows, Mac OS and Linux. One or multiple protein sequences in FASTA format could be inputted for predicting S-palmitoylation sites. Three predefined thresholds including 'high' (0.8920), 'medium' (0.7766) and 'low' (0.6484) were selected based on Sp values of ~95%, ~90% and ~85%, respectively (Supplementary Table S5). We also added an 'All'

option to enable the outputting of the prediction scores of all cysteine residues. The high threshold was adopted as the default cut-off value, and a manual was carefully written for users. GPS-Palm could be downloadable for academic research at http://gpspalm.biocuckoo.cn/download.php.

Results

A brief review of applicable methods for the prediction of S-palmitoylation sites

In 1979, Schmidt *et al*. [41] first discovered that palmitic acids could be covalently attached to Sindbis virus E2 and E1 glycoproteins and suggested that palmitoylation might play a potential role in regulating the maturation of viral proteins glycoproteins. Even after over 25 years, no well-defined consensus motifs were established from traditional experimental efforts [3, 4, 7, 42]. Although a considerable proportion of known S-palmitoylation sites follow the sequence pattern –CC– or –CXXC–, such simple motifs were difficult to enable an accurate site prediction [3, 4, 7, 42].

Advances in bioinformatics provided a great opportunity for the *in silico* prediction of PTM sites, and currently there have been 11 computational methods designed for predicting S-palmitoylation sites (Supplementary Table S1) [20, 21, 23, 24, 26–30]. In April 2006, we used GPS 1.0 algorithm, which was also called as clustering and scoring strategy (CSS) at that time, to develop the first tool named CSS-Palm 1.0 for the prediction of S-palmitoylation sites from protein sequences [20, 40]. Later, we improved the GPS algorithm and released CSS-Palm 2.0 [22], which was further updated into GPS-Lipid for an extended prediction of four types of lipid modifications besides S-palmitoylation [29]. We also used the Naïve Bayes algorithm to develop an alternative tool of NBA-Palm [21]. Besides our studies, other scientists also took great efforts in this field. In 2009, Wang *et al*. [23] encoded the CKSAAP feature and constructed a highly useful predictor CKSAAP-Palm, which was implemented with the support vector machine (SVM) algorithm. Using the same SVM algorithm, Li *et al*. [24] released PPWMs by integrating three types of sequence-based and structural features, including PSSM, ASA and SS. By developing IFS-Palm with the *k*-nearest neighbor (KNN) algorithm, Hu *et al*. [25] not only adopted two types of frequently used sequence features as AAindex and PSSM, but also considered intrinsically disordered regions (IDRs) in protein sequences, as well as three types of specific features including the distance to transmembrane domains, the distance to the N- or C-terminus of the protein, and a N-terminal MGC motif for S-palmitoylation sites. In 2013, Shi *et al*. [26] combined four types of features as ASA, PseAAC, ACF and PSSM, and three algorithms of KNN, SVM and decision tree (DT) were tested to construct an online service of WAP-Palm. Later, a SVM-based web server of PalmPred was released by incorporating PSSM, IDRs and SS features [27], whereas a random forest (RF)-based tool of SeqPalm was designed by encoding PseAAC, ACF and PSSM features [28]. In Jun 2017, by using the algorithms of maximal dependence decomposition (MDD) and SVM, Weng *et al*. [30] reported a high-quality predictor named MDD-Palm, in which up to five types of features were incorporated, including PseAAC, CKSAAP, PSSM, GPS and ASA. More details on these programs, including sizes of training data sets, data sources, features, algorithms, web links and window sizes for encoding PSP(*m*, *n*) items were shown, as well as original references (Supplementary Table S1).

For predicting S-palmitoylation sites, there were still three problems needed to be addressed. First, since the functional importance of S-palmitoylation has been more and more recognized by biologists, thousands of S-palmitoylation sites have been identified from both small- and large-scale studies in recent years. A larger training data set will benefit to develop a more accurate predictor. Second, various sequence-based and structural features were proposed, while it is not known whether they are all efficient in predicting S-palmitoylation sites based on an expanded data set. Finally, the existing tools were mainly implemented in traditionally machine-learning algorithms, which are less effective in feature extraction and presentation. Using deep learning algorithms might tackle this problem for a more accurate prediction.

Compilation of a large data set of known S-palmitoylation sites

Through the literature biocuration and public database integration, we obtained 3098 no-redundant S-palmitoylation sites in 1618 known substrates, after homology elimination (Supplementary Table S2). Compared with the data sets prepared in other studies, our benchmark data set was much larger, with a >4.2-fold increase of known sites (Figure 3A). In GPS-Lipid [29] and MDD-Palm [30], only 737 and 710 unique S-palmitoylation sites were collected, respectively (Figure 3A, Supplementary Table S1). In our data set, known S-palmitoylated proteins with corresponding sites were collected from 79 eukaryotic and prokaryotic organisms, including *Homo sapiens*, *Bos taurus*, *Oryctolagus cuniculus*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Escherichia coli*, *Mycobacterium tuberculosis* and other species (Figure 3B). Two species with the most of sites were *M. musculus* and *H. sapiens*, in which there were 1958 and 761 unique sites in 1049 and 354 proteins, respectively (Figure 3B).

In the benchmark data set, there were 1539 small-scale sites of 667 proteins verified by conventional experimental assays, whereas 1559 large-scale sites in 965 substrates were exclusively identified by MS (Supplementary Table S2). The distribution of proteins with different numbers of S-palmitoylation sites was counted, and we found that up to 55.17% of modified proteins contained ≥2 small-scale sites, while this proportion was decreased to 32.54% for large-scale sites (Figure 3C). We further counted the proportion of positive sites and observed that S-palmitoylation sites occupied 18.58 and 11.16% of all cysteine residues in S-palmitoylated proteins derived from the small- and large-scale studies, respectively (Figure 3D). Thus, the large-scale substrates tended to have fewer S-palmitoylation sites than the small-scale counterparts, and the results indicated that a considerable number of *bona fide* S-palmitoylation sites might be missed as false negatives by large-scale identification, probably due to the sensitivity limitation of MS instruments. Furthermore, we used pLogo (https://plogo.uconn.edu/) [43], a sequence logo generator to analyze amino acid preferences around the small- and large-scale sites (Figure 3E). For the small-scale sites, cysteine residues were enriched at positions from −5 to +5, and the two known S-palmitoylation motifs of –CC– and –CXXC– [3, 4, 7, 42] could fit such a sequence pattern well. On the contrary, the sequence profile of large-scale sites was quite elusive, and cysteine residues were not statistically over-represented in any positions of PSP(10, 10) items. In this regard, false positives might occur in large-scale studies. Taken together, our results demonstrated that the data quality of large-scale sites might be considerably lower than the small-scale sites.

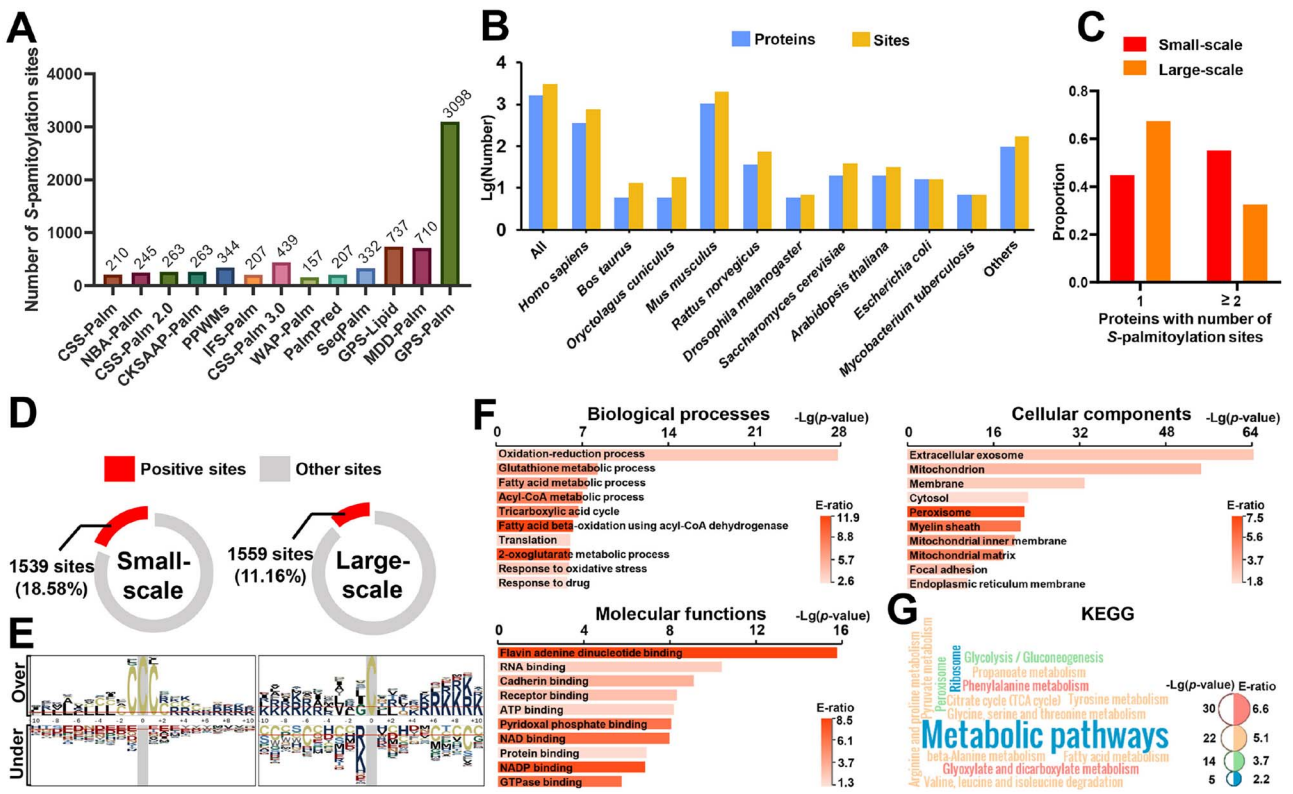

Figure 3. The analysis of the benchmark data set. (A) A comparison of the data set for GPS-Palm against other existing predictors. (B) The distribution of S-palmitoylated proteins and sites in several major organisms. (C) The distribution of proteins with different numbers of S-palmitoylation sites in small- or large-scale data set. (D) The proportion of positive sites in small- and large-scale data sets. (E) The sequence logos of small- and large-scale sites. Upper and lower characters denoted over- and under-represented amino acid residues in positive data sets at individual positions, respectively. The height of a residue was proportional to its statistical significance (log-odds of the binomial probability). (F) Enrichment analyses of mouse S-palmitoylated proteins based on GO biological processes, molecular functions and cellular components (P-value $< 10^{-5}$). (G) KEGG-based enrichment results (P-value $< 10^{-5}$).

Using the 1049 known S-palmitoylated proteins in *M. musculus* (Supplementary Table S2), we conducted an enrichment analysis based on gene ontology (GO) annotations [44], with the hypergeometric test. We observed that mostly enriched biological processes were metabolism-associated processes, such as oxidation–reduction process (GO:0055114), glutathione metabolic process (GO:0006749), fatty acid metabolic process (GO:0006631), acyl-CoA metabolic process (GO:0006637) and tricarboxylic acid cycle (GO:0006099) (Figure 3F). The results were highly consistent with previous studies, which demonstrated that S-palmitoylation plays a critical role in regulating cellular metabolism [45, 46]. Enrichment results of GO molecular functions and cellular components also indicated that mouse S-palmitoylated proteins were significantly over-represented in a variety of membrane-bound organelles and might be functional through interacting with multiple types of biomolecules (Figure 3F). Further analyses of the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [47, 48] supported the enrichment of S-palmitoylated proteins in metabolism pathways (Figure 3G). The GO- and KEGG-based enrichment analyses were also performed for 354 human S-palmitoylated proteins. Due to the data limitation, over-represented human biological processes and pathways were much more diverse than in *M. musculus* (Supplementary Figure S1A and B). However, a number of enriched GO cellular components and molecular functions, such as plasma membrane (GO:0005886), focal adhesion (GO:0005925), membrane raft (GO:0045121), integrin binding (GO:0005178) and cadherin binding (GO:0045296), supported human S-palmitoylation to be highly involved in membrane-associated functions (Supplementary Figure S1A). Due to the conservedness of S-palmitoylation regulation, the enrichment results would be helpful for further analyzing regulatory roles of S-palmitoylation in eukaryotes.

The heterogeneous data quality of small- and large-scale sites

Since both false negatives and false positives existed in large-scale sites, directly including these sites into the training data set might influence the prediction accuracy. To probe this problem, we used a simple but efficient feature named OBC [21] to encode PSP(10, 10) items, and the improved PLR algorithm was adopted for training models (Figure 4A, Supplementary Methods). We found the model trained on the small-scale sites achieved a higher AUC value of 0.718 from the 10-fold cross-validations (Figure 4B). However, mixing small- and large-scale sites together significantly reduced the 10-fold cross-validation AUC value to 0.684 (Two-tailed t-test, P-value $< 10^{-15}$), and exclusively using large-scale sites only produced an AUC score of 0.642 (Figure 4B). Thus, our results demonstrated that large-scale sites were highly error-prone and could not be equally treated as small-scale sites.

Since the data quality of small- and large-scale sites was obviously different, such a difference should be quantitatively measurable to discriminate the two types of S-palmitoylation sites. From the strong sequence pattern of small-scale sites

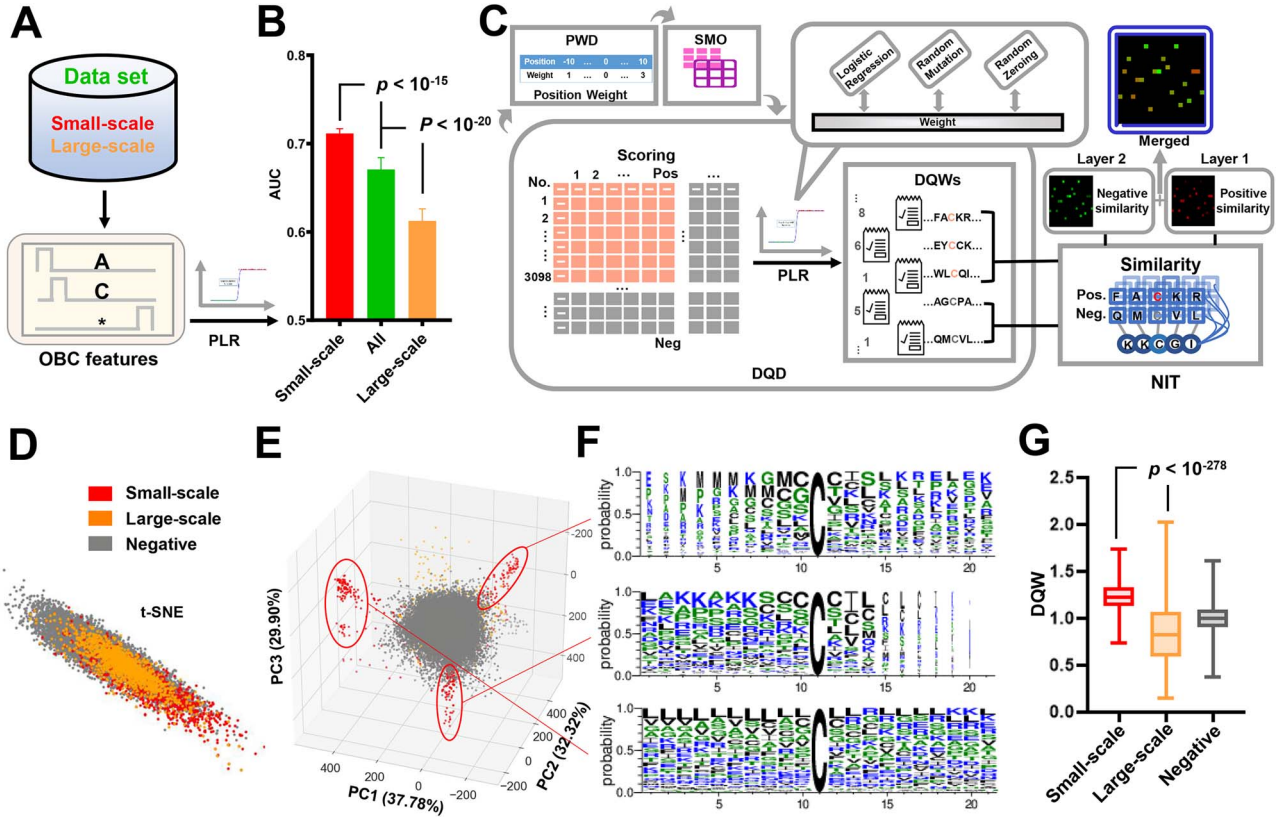

Figure 4. Development of DQD and NIT to distinguish small- and large-scale sites. (A) The OBC feature was adopted for encoding PSP(10, 10) items of small- and large-scale sites, respectively. The improved PLR algorithm was used for training. (B) The 10-fold cross-validation AUC values for models trained on small-scale, all positive and large-scale sites. (C) The implementation of DQD and NIT that transformed numerical similarity values to two-layer color images. (D) The *t*-SNE analysis of images generated from small-scale, large-scale and negative sites. (E) The PCA analysis of PSP(10, 10)-based images. (F) The sequence logos of Cluster I (N-terminal), II (C-terminal) and III (Internal) sites. The thinner characters in a column represented that a smaller number of amino acids appeared in the position. (G) Distribution of optimized DQWs for small-scale, large-scale and negative sites.

(Figure 3E), we assumed that *bona fide* *S*-palmitoylation sites should be more like each other, and nonpalmitoylated residues will also be all alike. Based on this hypothesis, we developed a new method named DQD, which automatically assigned a unique DQW for each PSP(10, 10) item in the positive and negative data sets, based on the optimal 10-fold cross-validation AUC value derived from the PLR-based training (Figure 4C). To enable the use of CNNs for model training, we designed an additional approach named NIT, which separately represented the similarity values of a given PSP(10, 10) item A against the positive and negative data sets into two matrices $Mat_+(A)$ and $Mat_-(A)$. Then, NIT transformed the two similarity matrices into a two-layer RGB image of 21×20 pixels (Figure 4C).

Using DQD and NIT, all small-scale, large-scale and negative sites in the benchmark data sets were transformed into individual images, which were further analyzed by *t*-distributed stochastic neighbor embedding (*t*-SNE). From the results, we found that the three types of sites could be clearly distinguished (Figure 4D). Most of the small-scale sites were far from negative sites, and large-scale sites were in between (Figure 4D). Furthermore, the principal component analysis (PCA) method was used to analyze the image data. It could be found that negative sites condensed a single cluster, and large-scale sites were distributed between small-scale and negative sites (Figure 4E). Interestingly, small-scale sites were separated into three distinct clusters (Figure 4E). Again, pLogo [43] was used to analyze the sequence preference for each cluster. Cluster I and II sites were palmitoylated at positions nearby N- and C-termini of protein sequences, whereas the Cluster III sites located in internal positions of proteins (Figure 4F). The distribution of optimized DQWs was analyzed, and we found that the average DQW value (0.824) of large-scale sites was significantly lower than small-scale (1.229) sites (Figure 4G).

Development of GPS-Palm for predicting S-palmitoylation sites

For the prediction of general *S*-palmitoylation sites, we encoded PSP(10, 10) items by using seven sequence-based features including GPS, PseAAC, CKSAAP, OBC, AAindex, ACF and PSSM, and three structural features including ASA, SS and BTA [20–30, 32] (Figure 1C, Supplementary Methods, and Supplementary Table S3). Then, pCNNs were adopted for training models (Figure 2), and predefined parameters in each network were present (Supplementary Table S4). The 10-fold cross-validation was performed for each feature, and AUC values ranged from 0.562 (BTA) to 0.806 (GPS 6.0) (Figure 5A). From the results, it could be found that the GPS feature was more informative than other features. By integrating the 10 types of features, GPS-Palm reached a 10-fold cross-validation AUC value as 0.855 (Figure 5A). For the GPS 6.0 algorithm, we further evaluated whether the two new methods of DQD and NIT-based CNNs could improve the prediction performance (Figure 5B). Through the 10-fold

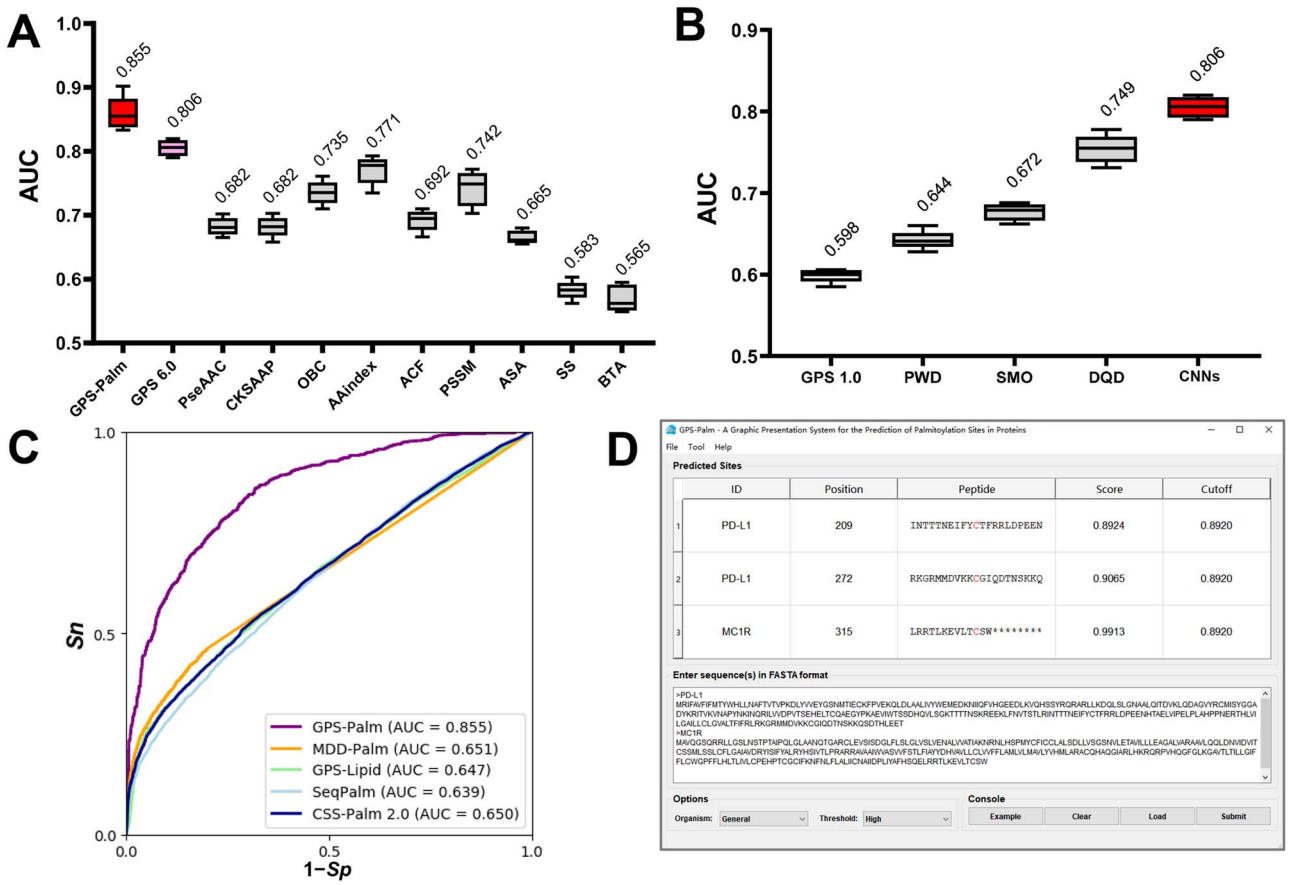

Figure 5. The development of GPS-Palm. (A) The 10-fold cross-validation AUC values were individually calculated for the 10 types of features. GPS-Palm integrated all these features and achieved a better accuracy. (B) In GPS 6.0, all methods including PWD, SMO, DQD and NIT-based CNNs were useful for performance improvement. (C) A comparison of GPS-Palm to four existing predictors, including CSS-Palm 2.0 [22], SeqPalm [28], GPS-Lipid [29] and MDD-Palm [30]. (D) The GUI interface of GPS-Palm. Two oncogenic proteins including PD-L1 [11] and MC1R [12] were selected as examples for the prediction of S-palmitoylation sites.

cross-validations, the original GPS 1.0 algorithm achieved an AUC value of 0.598, while the AUC score of GPS 5.0 algorithm with PWD and SMO could be increased to 0.672 (Figure 5B). Sequentially including DQD and CNNs could further enhance the accuracy to 0.749 and 0.806, respectively (Figure 5B).

Furthermore, we implemented two additional species-specific predictors by exclusively using human or mouse S-palmitoylation sites for model training (Supplementary Figure S2). For predicting human-specific sites, the 10-fold cross-validation AUC scores ranged from 0.665 (CKSAAP) to 0.865 (GPS 6.0) for individual features, whereas the pCNN-based integration of the 10 types of features achieved a better accuracy of 0.900 (Supplementary Figure S2A). Compared with GPS 1.0, sequentially adding PWD, SMO, DQD and NIT-based CNNs increased the AUC values from 0.616 to 0.643, 0.676, 0.754 and 0.865, respectively (Supplementary Figure S2B). To predict mouse-specific sites, the 10-fold cross-validation AUC scores were calculated from 0.651 (CKSAAP) to 0.849 (GPS 6.0) for individual features, and the combination of the 10 types of features reached a superior AUC value of 0.897 (Supplementary Figure S2C). Again, we compared GPS 6.0 to 1.0 and observed that each of the four new methods contributed to performance improvement (Supplementary Figure S2D).

To exhibit the superiority of GPS-Palm, we compared it to other existing tools, including CSS-Palm 2.0 [22], SeqPalm [28], GPS-Lipid [29] and MDD-Palm [30] (Supplementary Table S1). We directly submitted the benchmark data set into these tools to calculate the AUC values, which were compared with the 10-fold cross-validation result of GPS-Palm. It could be found that the accuracy of GPS-Palm was much higher than the second one, MDD-Palm [30], with a >31.34% increase of the AUC value (0.855 versus 0.651) (Figure 5C). Finally, local packages of GPS-Palm were implemented in Python 3.6 and PyQt 5.0, with an easy-to-use GUI interface (Supplementary Methods). Users could input one or multiple protein sequences, select a threshold and then click on 'Submit' for a prediction (Figure 5D). The results will be shown in a tabular format, including sequence ID, palmitoylated position, flanking peptide, predicted score and predefined cut-off value (Figure 5D). In addition, the button 'Load' could be clicked to load a sequence file for a large-scale prediction.

Discussion

As the only type of reversible lipid modification, S-palmitoylation plays a vital role in regulating a broad spectrum of biological processes [1–3, 7–11], whereas the dysregulation of protein S-palmitoylation is associated with human diseases [13, 14, 49]. For example, multiple cysteine residues in the nucleotide oligomerization domain (NOD)-like receptors 1 and 2 (NOD1/2) were discovered to be modified by the zinc finger DHHC-type palmitoyltransferase 5 (ZDHHC5), and S-palmitoylation of the two intracellular pattern-recognition proteins is essential for activating immune responses in bacterial sensing [50].

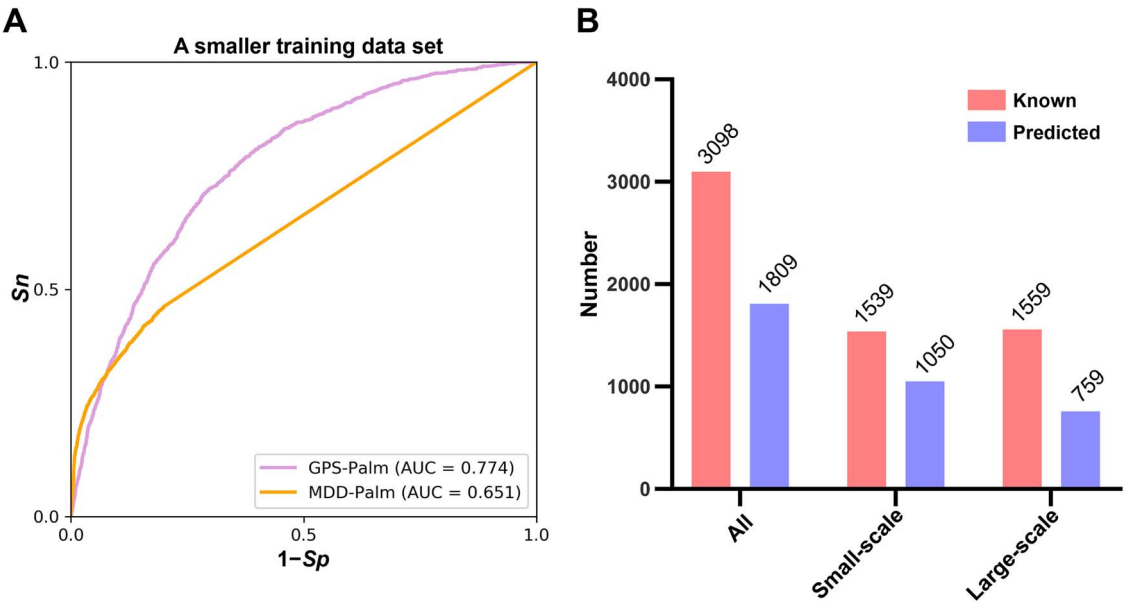

Figure 6. Additional analyses of GPS-Palm. (A) An additional comparison of GPS-Palm with MDD-Palm [30], by using the same training data set (Supplementary Table S1). The self-consistency accuracy of MDD-Palm was calculated and compared to the 10-fold cross-validation AUC value of GPS-Palm. (B) We used GPS-Palm to generate scores for all collected S-palmitoylation sites (known), and the number of sites predicted under the high threshold was counted (predicted) (Supplementary Table S2).

Also, Chen *et al.* [21] used our NBA-Palm 1.0 to predict melanocortin-1 receptor (MC1R), an important G-protein-coupled receptor (GPCR) in regulating mammalian pigmentation, to be potentially modified at C78 and C315 [12]. C315 was validated as the major MC1R S-palmitoylation site that is catalyzed by the palmitoyltransferase ZDHHC13, whereas MC1R palmitoylation enhances pigmentation and cell cycle arrest to inhibit melanomagenesis [12]. In particular, the site was predicted as the only positive hit by GPS-Palm under the high threshold (Figure 5D). More recently, Niu *et al.* [51] identified signal transducer and activator of transcription 3 (STAT3), a multifaceted oncoprotein, to be S-palmitoylated by ZDHHC19 at C687 and C712. STAT3 S-palmitoylation increases its homodimerization and transcriptional activity, and plays a critical role in promoting inflammation and tumorigenesis. In this regard, the identification of new S-palmitoylated substrates with exact sites is the foundation of understanding the molecular mechanisms and regulatory roles of S-palmitoylation. In contrast with tedious and laborious experimental assays [3, 4, 7, 15], computational predictions of S-palmitoylation sites in proteins can greatly narrow down potential candidates for further experimental consideration. To date, 11 predictors have been constructed for this purpose (Supplementary Table S1).

In this study, we first developed the GPS 6.0 algorithm by implementing two new methods of DQD and NIT-based CNNs (Figure 1A and B). The original methods of PWD and SMO in GPS 5.0 were reserved, as well as the basic scoring strategy. Using DQD could increase the 10-fold cross-validate AUC value from 0.672 to 0.749, and NIT-based CNNs further improved the AUC value to 0.806 (Figure 5B). In our benchmark data set, small-scale sites could be regarded as high-quality positive data. However, both false negatives and false positives existed in large-scale sites, which could be expected to have a lower data quality. Indeed, we found that DQWs of large-scale sites were significantly lower than in small-scale sites, from the results of DQD (Figure 4G). The integration of these large-scale sites by DQD and NIT rather than simply including or discarding them not only maximized the size of the training data set, but also achieved a dramatically increased accuracy for the prediction of S-palmitoylation sites. Besides GPS 6.0, we further integrated nine additional features and implemented a deep learning framework of pCNNs for model training (Figure 1C). By comparison, GPS-Palm was much better than other existing tools (Figure 5C). Since other tools used much smaller training data sets, the higher accuracy of GPS-Palm might be attributed to the methodology or the larger training data set. To exhibit the superiority of our methods, we re-retrained a pCNN model by using the training data set of MDD-Palm [30], which prepared 710 positive and 5676 negative sites (Supplementary Table S1). The 10-fold cross-validation AUC value was calculated as 0.774, which was much better than MDD-Palm [30] (Figure 6A). Using GPS-Palm, we assigned prediction scores for all small- and large-scale sites (Supplementary Table S2). From 3098 known S-palmitoylation sites, 1809 (58.4%) sites were predicted with scores greater than the high cut-off value of 0.8920 (Figure 6B). Under the high threshold, 68.2% (1050/1539) small-scale sites and 48.7% (759/1559) large-scale sites were predicted (Figure 6B). Thus, the data quality of small-scale sites was much higher, and our predictions will be useful to prioritize highly potential candidates for further experimental researches.

For the future, we will continue to maintain and improve GPS-Palm. The training data set will be enlarged when newly identified S-palmitoylation sites are available. Also, more sequence-based and structural features, as well as other types of cutting-edge artificial intelligence algorithms, will be tested and included if the prediction accuracy can be increased. Recently, we participated in a collaborative study, in which human programmed-death ligand 1 (PD-L1) was predicted and verified to be palmitoylated at C272, and this S-palmitoylation event inhibits T-cell-mediated immune responses against tumors, through blocking its mono-ubiquitination to prevent endosomal sorting complexes required for transport (ESCRT)-mediated sorting to the multivesicular body (MVB) and lysosomal degradation of PD-L1 [11]. Thus, the quantitative proteomic

and PTMomic data sets could be incorporated to predict functional impacts of S-palmitoylation sites. It should be noted that molecular fluctuations of big omic data could be simply visualized and represented in heatmap images, which can be directly processed by CNNs. Thus, it is possible to develop a more general presentation system to integrate both sequences and omic data sets into a single framework to achieve a higher accuracy for predicting PTM substrates and/or sites.

Taken together, although many efforts can be taken in the near future, our study provided a highly accurate tool for predicting S-palmitoylation sites from protein sequences. We anticipate that the methods used in this work can be easily extended for other types of PTM predictions.

Key Points

- We reviewed the 11 existing tools for the prediction of S-palmitoylation sites, while the training data sets, features and algorithms used in these tools were summarized.
- We developed a new method named data quality discrimination (DQD) to measure and discriminate different data quality weights (DQWs) of S-palmitoylation sites identified from small- or large-scale experiments.
- We encoded numerical features into images, integrated DQD and convolutional neural networks (CNNs) into our previous methods and developed a new algorithm of graphic presentation system (GPS) 6.0.
- We further incorporated nine additional features and implemented a framework of parallel CNNs (pCNNs) to develop a new tool of GPS-Palm, which exhibited a higher accuracy than other exiting tools.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

Funding for open access charge: Special Project on Precision Medicine under the National Key R&D Program (2017YFC0906600, 2018YFC0910500), the Natural Science Foundation of China (31930021, 31970633, 31671360, 81701567), the Fundamental Research Funds for the Central Universities (2017KFXKJC001, 2019kfyRCPY043), Changjiang Scholars Program of China and the program for HUST Academic Frontier Youth Team.

References

1. Ray A, Jatana N, Thukral L. Lipidated proteins: spotlight on protein-membrane binding interfaces. *Prog Biophys Mol Biol* 2017;**128**:74–84.
2. Casey PJ. Protein lipidation in cell signaling. *Science* 1995;**268**:221–5.
3. Roth AF, Wan J, Bailey AO, *et al*. Global analysis of protein palmitoylation in yeast. *Cell* 2006;**125**:1003–13.
4. Dietrich LE, Ungermann C. On the mechanism of protein palmitoylation. *EMBO Rep* 2004;**5**:1053–7.
5. Greaves J, Chamberlain LH. Palmitoylation-dependent protein sorting. *J Cell Biol* 2007;**176**:249–54.
6. Linder ME, Deschenes RJ. Palmitoylation: policing protein stability and traffic. *Nat Rev Mol Cell Biol* 2007;**8**:74–84.
7. Smotrys JE, Linder ME. Palmitoylation of intracellular signaling proteins: regulation and function. *Annu Rev Biochem* 2004;**73**:559–87.
8. Kleuss C, Krause E. Galpha(s) is palmitoylated at the N-terminal glycine. *EMBO J* 2003;**22**:826–32.
9. Shen LF, Chen YJ, Liu KM, *et al*. Role of S-palmitoylation by ZDHHC13 in mitochondrial function and metabolism in liver. *Sci Rep* 2017;**7**:2182.
10. Kim SW, Kim DH, Park KS, *et al*. Palmitoylation controls trafficking of the intracellular Ca(2+) channel MCOLN3/TRPML3 to regulate autophagy. *Autophagy* 2019;**15**:327–40.
11. Yao H, Lan J, Li C, *et al*. Inhibiting PD-L1 palmitoylation enhances T-cell immune responses against tumours. *Nat Biomed Eng* 2019;**3**:306–17.
12. Chen S, Zhu B, Yin C, *et al*. Palmitoylation-dependent activation of MC1R prevents melanomagenesis. *Nature* 2017;**549**:399–403.
13. Andrew RJ, Fernandez CG, Stanley M, *et al*. Lack of BACE1 S-palmitoylation reduces amyloid burden and mitigates memory deficits in transgenic mouse models of Alzheimer's disease. *Proc Natl Acad Sci U S A* 2017;**114**:E9665–74.
14. Berchtold LA, Storling ZM, Ortis F, *et al*. Huntingtin-interacting protein 14 is a type 1 diabetes candidate protein regulating insulin secretion and beta-cell apoptosis. *Proc Natl Acad Sci U S A* 2011;**108**:E681–8.
15. Drisdel RC, Green WN. Labeling and quantifying sites of protein palmitoylation. *Biotechniques* 2004;**36**:276–85.
16. Martin BR, Cravatt BF. Large-scale profiling of protein palmitoylation in mammalian cells. *Nat Methods* 2009;**6**:135–8.
17. Yang W, Di Vizio D, Kirchner M, *et al*. Proteome scale characterization of human S-acylated proteins in lipid raft-enriched and non-raft membranes. *Mol Cell Proteomics* 2010;**9**:54–70.
18. Forrester MT, Hess DT, Thompson JW, *et al*. Site-specific analysis of protein S-acylation by resin-assisted capture. *J Lipid Res* 2011;**52**:393–8.
19. Collins MO, Woodley KT, Choudhary JS. Global, site-specific analysis of neuronal protein S-acylation. *Sci Rep* 2017;**7**:4683.
20. Zhou F, Xue Y, Yao X, *et al*. CSS-palm: palmitoylation site prediction with a clustering and scoring strategy (CSS). *Bioinformatics* 2006;**22**:894–6.
21. Xue Y, Chen H, Jin C, *et al*. NBA-palm: prediction of palmitoylation site implemented in naive Bayes algorithm. *BMC Bioinformatics* 2006;**7**:458.
22. Ren J, Wen L, Gao X, *et al*. CSS-palm 2.0: an updated software for palmitoylation sites prediction. *Protein Eng Des Sel* 2008;**21**:639–44.
23. Wang XB, Wu LY, Wang YC, *et al*. Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs. *Protein Eng Des Sel* 2009;**22**:707–12.
24. Li YX, Shao YH, Deng NY. Improved prediction of palmitoylation sites using PWMs and SVM. *Protein Pept Lett* 2011;**18**:186–93.
25. Hu LL, Wan SB, Niu S, *et al*. Prediction and analysis of protein palmitoylation sites. *Biochimie* 2011;**93**:489–96.
26. Shi SP, Sun XY, Qiu JD, *et al*. The prediction of palmitoylation site locations using a multiple feature extraction method. *J Mol Graph Model* 2013;**40**:125–30.

27. Kumari B, Kumar R, Kumar M. PalmPred: an SVM based palmitoylation prediction method using sequence profile information. *PLoS One* 2014;**9**:e89246.
28. Li S, Li J, Ning L, *et al*. In Silico identification of protein S-palmitoylation sites and their involvement in human inherited disease. *J Chem Inf Model* 2015;**55**:2015–25.
29. Xie Y, Zheng Y, Li H, *et al*. GPS-lipid: a robust tool for the prediction of multiple lipid modification sites. *Sci Rep* 2016;**6**:28249.
30. Weng SL, Kao HJ, Huang CH, *et al*. MDD-palm: identification of protein S-palmitoylation sites with substrate motifs based on maximal dependence decomposition. *PLoS One* 2017;**12**:e0179529.
31. Poplin R, Chang PC, Alexander D, *et al*. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* 2018;**36**:983–7.
32. Yang Y, Heffernan R, Paliwal K, *et al*. SPIDER2: a package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. *Methods Mol Biol* 2017;**1484**:55–63.
33. Huang KY, Su MG, Kao HJ, *et al*. dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic Acids Res* 2016;**44**:D435–46.
34. Blanc M, David F, Abrami L, *et al*. SwissPalm: protein palmitoylation database. *F1000Res* 2015;**4**:261.
35. Keshava Prasad TS, Goel R, Kandasamy K, *et al*. Human protein reference database—2009 update. *Nucleic Acids Res* 2009;**37**:D767–72.
36. Xu H, Wang Y, Lin S, *et al*. PTMD: a database of human disease-associated post-translational modifications. *Genomics Proteomics Bioinformatics* 2018;**16**:244–51.
37. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017;**45**:D158–69.
38. Huang Y, Niu B, Gao Y, *et al*. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010;**26**:680–2.

39. Zhou FF, Xue Y, Chen GL, *et al*. GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem Biophys Res Commun* 2004;**325**:1443–8.
40. Xue Y, Ren J, Gao X, *et al*. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol Cell Proteomics* 2008;**7**:1598–608.
41. Schmidt MF, Bracha M, Schlesinger MJ. Evidence for covalent attachment of fatty acids to Sindbis virus glycoproteins. *Proc Natl Acad Sci U S A* 1979;**76**:1687–91.
42. Bijlmakers MJ, Marsh M. The on-off story of protein palmitoylation. *Trends Cell Biol* 2003;**13**:32–42.
43. O'Shea JP, Chou MF, Quader SA, *et al*. pLogo: a probabilistic approach to visualizing sequence motifs. *Nat Methods* 2013;**10**:1211–2.
44. Gene Ontology Consortium. The gene ontology (GO) project in 2006. *Nucleic Acids Res* 2006;**34**:D322–6.
45. Yang Q, Vijayakumar A, Kahn BB. Metabolites as regulators of insulin sensitivity and metabolism. *Nat Rev Mol Cell Biol* 2018;**19**:654–72.
46. Ko PJ, Dixon SJ. Protein palmitoylation and cancer. *EMBO Rep* 2018;19:pii: e46666.
47. Ogata H, Goto S, Sato K, *et al*. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 1999;**27**:29–34.
48. Ning W, Lin S, Zhou J, *et al*. WocEA: the visualization of functional enrichment results in word clouds. *J Genet Genomics* 2018;**45**:415–7.
49. Saleem AN, Chen YH, Baek HJ, *et al*. Mice with alopecia, osteoporosis, and systemic amyloidosis due to mutation in Zdhhc13, a gene coding for palmitoyl acyltransferase. *PLoS Genet* 2010;**6**:e1000985.
50. Lu Y, Zheng Y, Coyaud É, *et al*. Palmitoylation of NOD1 and NOD2 is required for bacterial sensing. *Science* 2019;**366**:460–7.
51. Niu J, Sun Y, Chen B, *et al*. Fatty acids and cancer-amplified ZDHHC19 promote STAT3 activation through S-palmitoylation. *Science* 2019;**573**:139–43.