Defining upstream enhancing and inhibiting sequence patterns for plant peroxisome targeting signal type 1 using large-scale *in silico* and *in vivo* analyses

Qianwen Deng[1,2,†], He Li[3,†], Yanlei Feng[2,†], Ruonan Xu[1], Weiran Li[1], Rui Zhu[1], Delara Akhter[1,4], Xingxing Shen[1] (iD), Jianping Hu[5,*] (iD), Hangjin Jiang[3,*] (iD) and Ronghui Pan[1,2,*] (iD)

[1]College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China,
[2]ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou 310027, China,
[3]Center for Data Science, Zhejiang University, Hangzhou 310058, China,
[4]Department of Genetics and Plant Breeding, Sylhet Agricultural University, Sylhet 3100, Bangladesh, and
[5]Department of Energy Plant Research Laboratory and Plant Biology Department, Michigan State University, East Lansing, Michigan 48824, USA

SUMMARY

Peroxisomes are universal eukaryotic organelles essential to plants and animals. Most peroxisomal matrix proteins carry peroxisome targeting signal type 1 (PTS1), a C-terminal tripeptide. Studies from various kingdoms have revealed influences from sequence upstream of the tripeptide on peroxisome targeting, supporting the view that positive charges in the upstream region are the major enhancing elements. However, a systematic approach to better define the upstream elements influencing PTS1 targeting capability is needed. Here, we used protein sequences from 177 plant genomes to perform large-scale and in-depth analysis of the PTS1 domain, which includes the PTS1 tripeptide and upstream sequence elements. We identified and verified 12 low-frequency PTS1 tripeptides and revealed upstream enhancing and inhibiting sequence patterns for peroxisome targeting, which were subsequently validated *in vivo*. Follow-up analysis revealed that nonpolar and acidic residues have relatively strong enhancing and inhibiting effects, respectively, on peroxisome targeting. However, in contrast to the previous understanding, positive charges alone do not show the anticipated enhancing effect and that both the position and property of the residues within these patterns are important for peroxisome targeting. We further demonstrated that the three residues immediately upstream of the tripeptide are the core influencers, with a 'basic-nonpolar-basic' pattern serving as a strong and universal enhancing pattern for peroxisome targeting. These findings have significantly advanced our knowledge of the PTS1 domain in plants and likely other eukaryotic species as well. The principles and strategies employed in the present study may also be applied to deciphering auxiliary targeting signals for other organelles.

Keywords: amino acid polarity and charge, large-scale statistical analysis, organelles, peroxisome targeting signal type 1 (PTS1), protein subcellular localization, upstream enhancing and inhibiting patterns.

INTRODUCTION

Peroxisomes are universal eukaryotic organelles housing various metabolic pathways and are functionally connected with other organelles such as mitochondria, chloroplasts, lipid bodies, and the endoplasmic reticulum. Severe peroxisomal dysfunction can cause fatal human genetic disorders and plant embryonic lethality (Honsho et al., 2020; Hu et al., 2012). The proteome and metabolism of peroxisomes vary significantly among different organisms, tissue types, and developmental stages, as well as in response to various environmental conditions (Corpas, 2019; Gabaldón, 2010; Pan et al., 2020; Reumann & Bartel, 2016). To completely understand the function and dynamics of peroxisomes, it is essential to better understand how peroxisomal proteins are targeted to these organelles.

The subcellular localization of a protein is largely driven by its targeting peptides, which differ in structure, length, and position for targeting to different organelles. Most

proteins destined for chloroplasts, mitochondria, and the secretory pathway use N-terminal targeting peptides (Chu et al., 2020; Emanuelsson et al., 2007; Murcha et al., 2014; Teufel et al., 2022). Peroxisomal matrix proteins, on the other hand, rely on two types of peroxisomal targeting signals (PTS) located at the C- (PTS1) and N-terminus (PTS2), respectively. PTS1, which is carried by most peroxisomal matrix proteins at the extreme C-terminus, was initially recognized as a tripeptide with a 'canonical' consensus of [S/A]-[K/R]-[L/M]. However, more and more 'non-canonical' derivatives have been discovered, demonstrating the complexity of PTS1 and the likely existence of many unknown PTS1 tripeptides (Brocard & Hartig, 2006; Lametschwandtner et al., 1998; Lingner et al., 2011; Reumann & Chowdhary, 2018).

A dilemma in protein subcellular localization is that some proteins with the exact same targeting peptides can have varied levels of targeting efficiency or even lack of targeting to a particular organelle. This indicates that factors other than the targeting peptides also influence protein targeting. For example, the targeting peptides may be masked by other regions of the same protein as a result of protein folding or by other interacting proteins. They may also be impeded by other structural barriers such as the targeting peptides for other organelles or transmembrane domains that inhibit cross-membrane transport. Furthermore, the function of targeting peptides may be compromised by the presence of inhibiting elements or the absence of enhancing elements nearby, which may be more important for weak targeting peptides that rely on auxiliary targeting elements. Studies in plants, yeasts, and animals led to the conclusion that basic residues with positive charges in the upstream sequence of the PTS1 tripeptides can enhance the peroxisome targeting ability of PTS1 (Bongcam et al., 2000; Chowdhary et al., 2012; Distel et al., 1992; Kragler et al., 1998; Lametschwandtner et al., 1998; Ma & Reumann, 2008; Neuberger et al., 2003; Reumann, 2004). However, most of these studies used relatively small-sized *in silico* data sets and lacked systematic validations. Questions concerning how essential the upstream positive charges are to different PTS1 tripeptides and at which upstream positions the positively charged residues function effectively remain unanswered. Furthermore, sequence patterns upstream of the PTS1 tripeptide that enhance or inhibit peroxisome targeting remain largely undefined. A systematic analysis of the PTS1 domain, which includes the tripeptide and its upstream sequence, is needed.

To dissect the PTS1 domain, we collected large data sets of peptide sequences from 177 higher plant genomes and used 9806 PTS1-containing peroxisomal proteins and 34 277 non-peroxisomal proteins for *in silico* analysis. After identifying 12 low-frequency plant PTS1 tripeptides, we retrieved peroxisomal proteins with rare-occurring

PTS1 tripeptides and non-peroxisomal proteins with established PTS1 tripeptides to deduce upstream enhancing and inhibiting sequence patterns. We verified these upstream sequence patterns and systematically elucidated the importance of residue positions and properties for targeting. Our discoveries significantly advanced the understanding of the role of upstream sequence in the peroxisome targeting capacity of various plant PTS1 peptides. Our approaches and findings can be applied to the investigation of peroxisome targeting in other eukaryotes. The strategies employed in our study may also be used to elucidate auxiliary targeting sequences for other organelles.

RESULTS

Assembly of large peroxisomal and non-peroxisomal protein data sets

To define the upstream enhancing or inhibitory sequence patterns of the PTS1 tripeptides, large databases for PTS1-containing peroxisomal and non-peroxisomal proteins were needed. For this purpose, we first generated a reference list of 81 PTS1-containing peroxisomal proteins (Figure 1) from known Arabidopsis peroxisomal proteins (Pan et al., 2018; Pan & Hu, 2018). Only proteins that have been experimentally validated to be peroxisomal or belong to a well-established peroxisomal pathway were included (Table S1). We found that, in Arabidopsis proteins containing both PTS1 and PTS2 signals, the PTS1 signal is often poorly conserved across plant species (Figure S1 and Table S2); therefore, all Arabidopsis proteins containing a putative PTS2 signal were excluded from the reference list. Using this reference list, we performed BLAST searches (https://blast.ncbi.nlm.nih.gov) for homologs in the genomes of 177 angiosperms, including 110 eudicots, 58 monocots, and nine basal species (Figure 1 and Table S1), and assembled a 'mutual best-match' data set of 9806 PTS1-containing peroxisomal proteins (Figure 1 and Table S1). To ensure the reliability of this data set, we took a further screening step (Figure 1) by filtering out the PTS1 tripeptides shared by fewer than 10 protein sequences (i.e. frequency of occurrence < 10) (Table S1). This procedure narrowed down the number of PTS1 tripeptides from 136 to 41, which included canonical sequences such as SKL>, SRL>, and AKL>, as well as non-canonical ones such as SLM>, SSM>, and SYI> (Table 1 and Table S1), where > indicates the stop codon.

To generate a database of non-peroxisomal proteins for references in a subsequent analysis of the PTS1 domain, we performed similar BLAST searches for homologs of 476 Arabidopsis transcription factors (TFs) (Figure 1), as TFs have not been found in the peroxisomal matrix. Using TFs from six previously well-characterized protein families, namely, MYB, basic leucine zipper (bZIP), auxin response factor (ARF), NAC (NAM, ATAF1/2, CUC2), WRKY, and
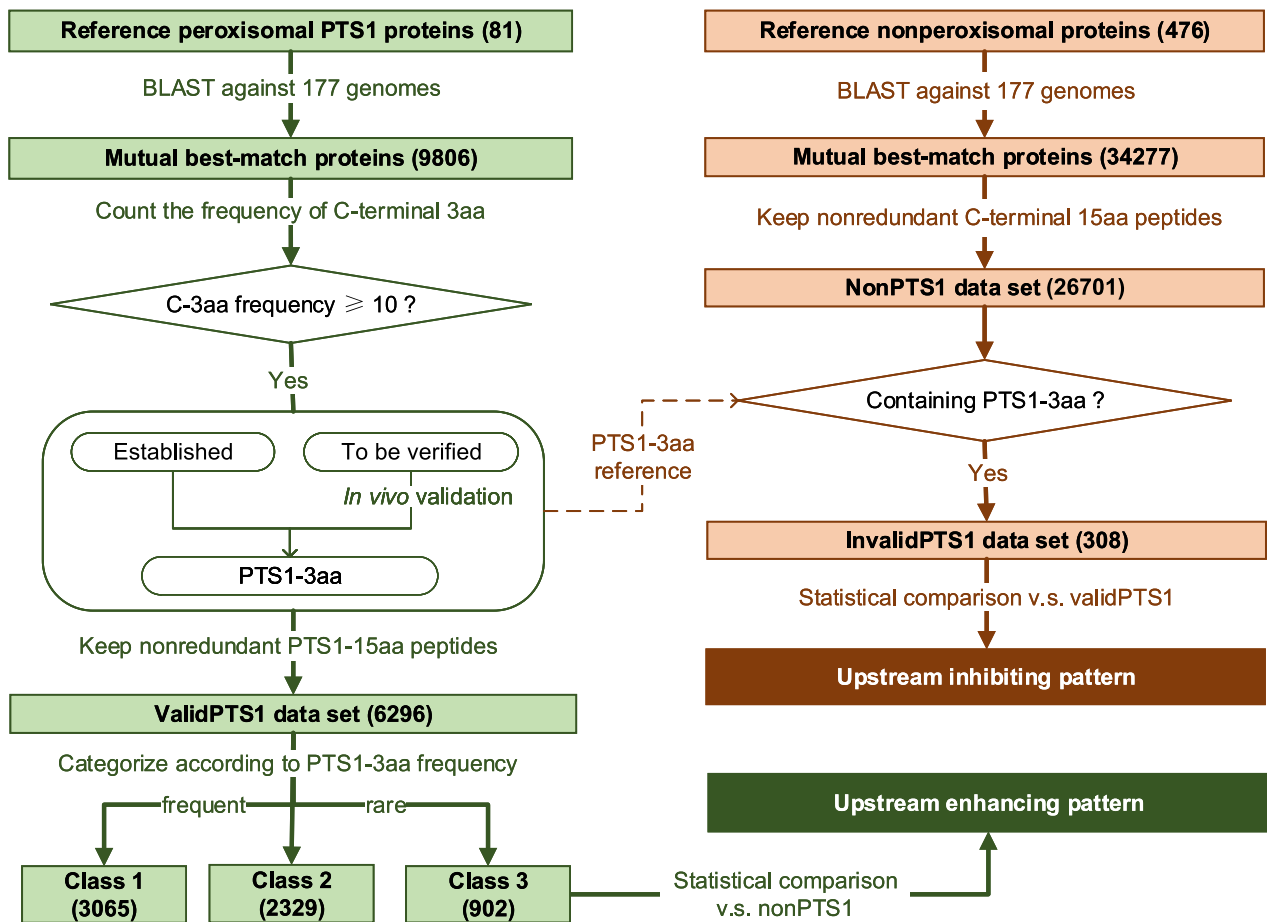
Figure 1. Workflow of data set generation in the present study.

PTS1-3aa indicates the PTS1 tripeptide. Numbers in parentheses indicate the number of protein or peptide sequences in the corresponding data set. Green and red boxes represent the data set generation processes for analyzing the upstream enhancing and inhibiting patterns, respectively. Diamonds indicate filtering conditions. Rounded squares indicate the verification process of PTS1-3aa.

Table 1 Plant PTS1 tripeptides identified in the present study

Previously established	Newly verified
SKL>, SRL>, AKL>, SRM>, SRI>, SSL>, SKM>, SKI>, ARL>, PRL>, SNL>, PKL>, SYM>, ASL>, SML>, SFM>, SNM>, CKL>, SGL>, SRV>, TKL>, AKI>, STL>, SAL>, SHL>, ALL>, SSM>, SLM>, PSL>	ARM>, SLL>, SNI>, SCI>, AKM>, ANL>, SFL>, SKF>, SYI>, SQL>, PRM>, CRL>

MADS (Table S3), a data set of 34 277 non-peroxisomal proteins was generated (Figure 1 and Table S3).

To dissect the PTS1 domain, we collected all the PTS1-15aa peptides from our peroxisomal PTS1 protein data set (Table S1). Previous studies reported the involvement of 12 or 14 amino acids at the C-terminus in peroxisome targeting (Emanuelsson et al., 2003; Lingner et al., 2011;

Reumann et al., 2012). Here, we chose the C-terminal 15-aa peptide (PTS1-15aa) to represent the PTS1 domain. After filtering out the redundant peptides, we generated a 'validPTS1 data set' that contained 6296 peptides (Table S4) to be used in subsequent analyses (Figure 1). As a reference, a 'nonPTS1 data set' of 26 701 peptides was also generated (Table S5), using all the non-redundant C-terminal 15-aa peptides from the non-peroxisomal protein data set (Figure 1 and Table S3).

Identification and verification of 12 low-frequency PTS1 tripeptides

Most of the 41 PTS1 tripeptides identified from the peroxisomal PTS1 protein data set (Table S1) have been found in known Arabidopsis PTS1-containing proteins or functionally established via subcellular localization analysis (Table 1) (Lingner et al., 2011; Ma & Reumann, 2008; Mullen et al., 1997; Pan et al., 2018; Pan & Hu, 2018; Ramirez et al., 2014). However, some tripeptides such as ARM>, SCI>,

SFL>, SYI>, and PRM> had never been experimentally validated in plants. Others such as SNI>, AKM>, CRL>, SLL>, ANL>, SKF>, and SQL> had been previously speculated to be peroxisomal or found to localize to punctate structures not verified to be peroxisomal (Table 1) (Lingner et al., 2011; Mullen et al., 1997). Interestingly, all these 12 tripeptides occurred at relatively low frequencies ($\leq$ 85) in our data set, in contrast to those of the canonical tripeptides SKL> (2871) and SRL> (1749) (Table S1).

We reasoned that some tripeptides with lower frequencies may have species specificity and therefore were not found in the known peroxisomal proteins. To test this possibility, we compared the frequency of each PTS1 tripeptide in different plant lineages, including the basal plant species, eudicots, and monocots, as well as in several representative model species such as Arabidopsis, soybean (*Glycine max*), rice (*Oryza sativa*), and maize (*Zea mays*) (Figure 2a). High-frequency PTS1 tripeptides, including [S/A]-[K/R]-[L/M]>, SRI>, SSL>, and SKI>, were conserved across plant lineages and species, whereas PTS1 tripeptides with lower frequencies were often only present in a few lineages or species (Figure 2a). These data suggested that infrequent PTS1 tripeptides tend to be species specific, underscoring the importance of including many species in the present study to uncover rare PTS1 tripeptides.

To validate these 12 low-frequency PTS1 tripeptides, we performed *in vivo* subcellular localization studies. We coinfiltrated Agrobacteria containing the mVenus-PTS1-15aa construct and those containing the peroxisome marker construct moxCerulean3-PTS1 (SKL) into tobacco leaves. Agrobacteria containing the mVenus-PTS1-15aa construct alone was used as a control to exclude the possible peroxisomal targeting of mVenus-PTS1-15aa by piggybacking onto moxCerulean3-PTS1, a mechanism known to exist in peroxisomes (Falter et al., 2019). Three days after infiltration, we used confocal fluorescence microscopy to observe the localization of the fluorescent proteins. Twelve of the 15 mVenus-PTS1-15aa fusions displayed complete peroxisome targeting, and the three fusion proteins containing SCI> showed both peroxisomal and cytosolic localizations, and presumably some diffusion into the nucleus (Figure 2b).

Taken together, we have identified and validated 12 low-frequency PTS1 tripeptides from the newly generated peroxisomal PTS1 protein data set. The finding that all the newly-verified PTS1 tripeptides were functional in the *in vivo* targeting analysis also validated the reliability of our data set.

### Correlation between the frequency of the PTS1 tripeptides and their upstream patterns

We posited that tripeptides with weaker targeting strength are more dependent on the upstream elements. In other words, functional PTS1 peptides with weaker tripeptides may be longer or more complex compared to those with strong tripeptides, and therefore more susceptible to random mutations within the tripeptide and/or upstream sequence. Consistent with this, almost every orthologous group of PTS1 proteins in our data set had one or a few dominant, strong tripeptides (Figure S2) that are expected to be evolutionarily more stable and thus more frequently occurring. We further hypothesized that different frequencies of the PTS1 tripeptides may be associated with the strength of their targeting ability.

To test this hypothesis, we first categorized all the PTS1-15aa peptides in the validPTS1 data set into three classes based on their frequency of occurrence (Figure 1 and Table S4). Class 1 included the PTS1-15aa peptides containing the two dominant and commonly used canonical tripeptides, SKL> (1837 samples) and SRL> (1228 samples), which together constituted 48.68% of the validPTS1 data set (Figure 3a). Class 2 included those containing seven relatively frequent tripeptides (i.e. AKL>, ARL>, SKM>, SRM>, SKI>, SRI>, and SSL>), which together were present in 2329 samples and counted for 36.99% of the data set (Figure 3a). Class 3 consisted of 32 low-frequency (< 200) tripeptides, which occurred in 902 (14.33%) of the samples (Figure 3a).

To analyze the relation between the frequency of the PTS1 tripeptides and their upstream signals, we calculated the Kullback–Leibler (KL) distance (or KL divergence) at every upstream position (−15 to −4) between each class of the validPTS1 data set and the nonPTS1 data set (Figure 3b). The KL distance value represents the relative entropy contained in each position, which indicates how different the experimental and the reference samples are at one position; for example, a KL value of 0 indicates no difference. The KL distances of class 1 and class 2 were similar, with class 2 slightly higher than class 1 at most positions (Figure 3b). By contrast, class 3 had an obviously higher KL compared to those of the first two classes in almost all the 12 positions (Figure 3b). Additionally, the three positions immediately upstream of the tripeptide (−6 to −4) showed relatively high KL distance values in all three classes, with class 1 being the lowest and class 3 the highest (Figure 3b).

Parallel to the KL analysis, we also performed seqlogo analysis of the PTS1-15aa peptides by comparing amino acid residues in each upstream position (Figure 3c). In a seqlogo, the *y*-axis represents the information content value where bigger-sized letters indicate lower variability, higher evolutionary conservation, and higher information content (Wagih, 2017). To select residues preferentially enriched in the upstream region of the valid PTS1-15aa peptides, information content value of the nonPTS1 data set was treated as the 'background noise' and thus was subtracted from the values of sequences from the validPTS1 data set. As expected, the 'strength' of the
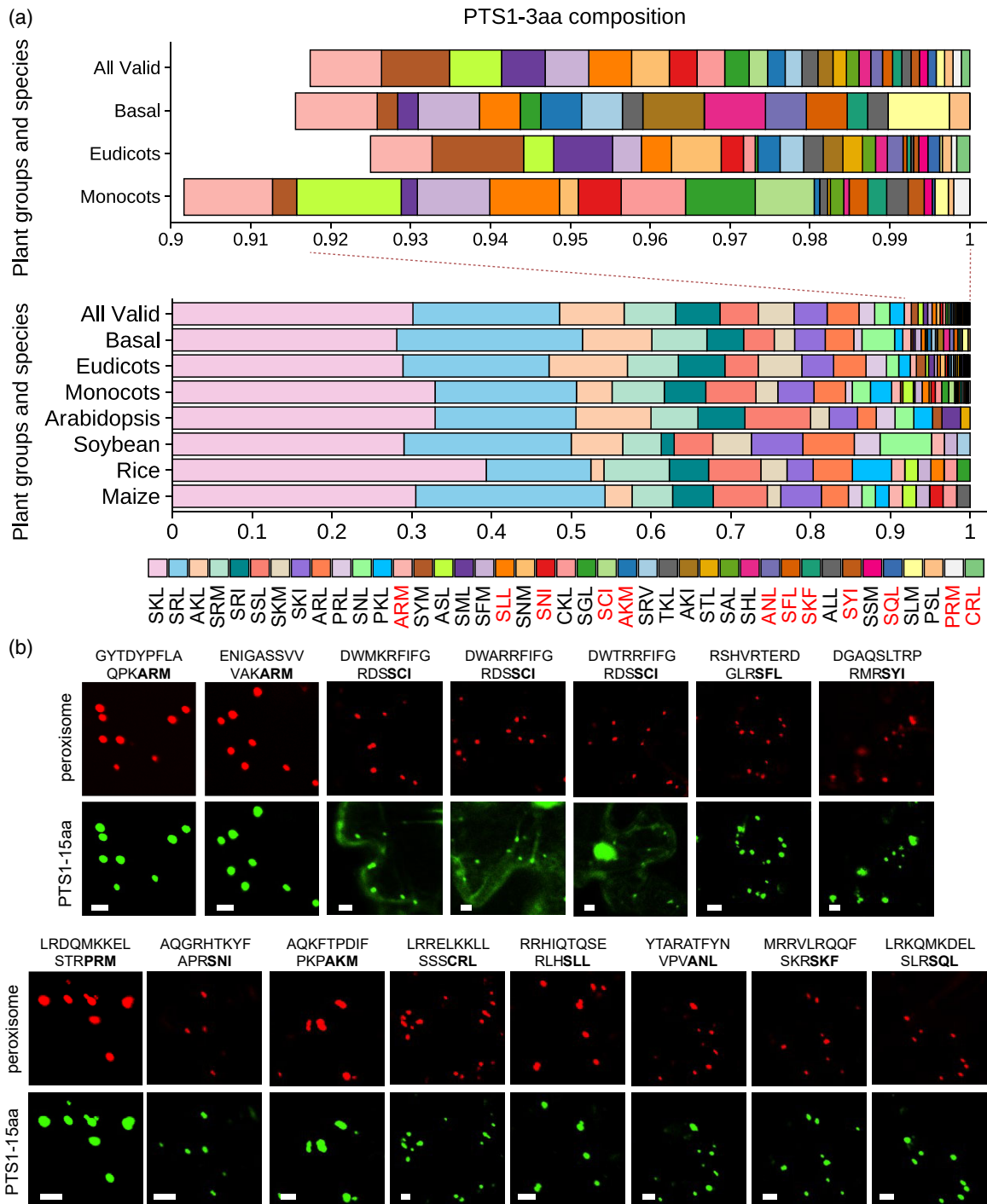
Figure 2. Analysis of the composition of PTS1 tripeptides in various plant genomes and verification of some of the low-frequency tripeptides.

(a) Distribution of various PTS1 tripeptides in plant genomes. The length of each colored bar correlates with the percentage of each PTS1 tripeptide within the plant group or species. The right ends of the bar graphs for four of the groups are also magnified for better visualization. All the tripeptides identified are listed at the bottom; the newly identified are in red.

(b) Confocal images of tobacco leaf cells transiently expressing the peroxisome marker moxCerulean3-PTS1 and the mVenus-PTS1-15aa peptide fusions containing the low-frequency tripeptides. Scale bars = 5 μm. Sources of the PTS1-15aa peptides: ENIGASSVVAKARM (*Eremochloa ophiuroides* OPR3), ENIGASSVVAKARM (*Arachis duranensis* ICL), DWTRRFIFGRDSSCI (*Digitaria exilis* NDB1), DWMKRFIFGRDSSCI (*Ananas comosus* NDB1), DWARRFIFGRDSSCI (*Dioscorea rotundata* NDB1), RSHVRTERDGLRSFL (*Cinnamomum micranthum* HAOX1), DGAQSLTRPRMRSYI (*Betula platyphylla* SDRc), LRDQMKKELSTRPRM (*Spirodela intermedia* AAE18), AQGRHTKYFAPRSNI (*Erysimum cheiranthoides* ST4), AQKFTPDIFPKPAKM (*Catharanthus roseus* SCP2), LRRELKKLLSSSCRL (*Gossypium hirsutum* 4CL13), RRHIQTQSERLHSLL (*Nelumbo nucifera* HAOX1), YTARATFYNVPVANL (*Aquilegia coerulea* ST3), MRRVLRQQFSKRSKF (*Musa schizocarpa* AAE17), and LRKQMKDELSLRSQL (*Benincasa hispida* AAE18).
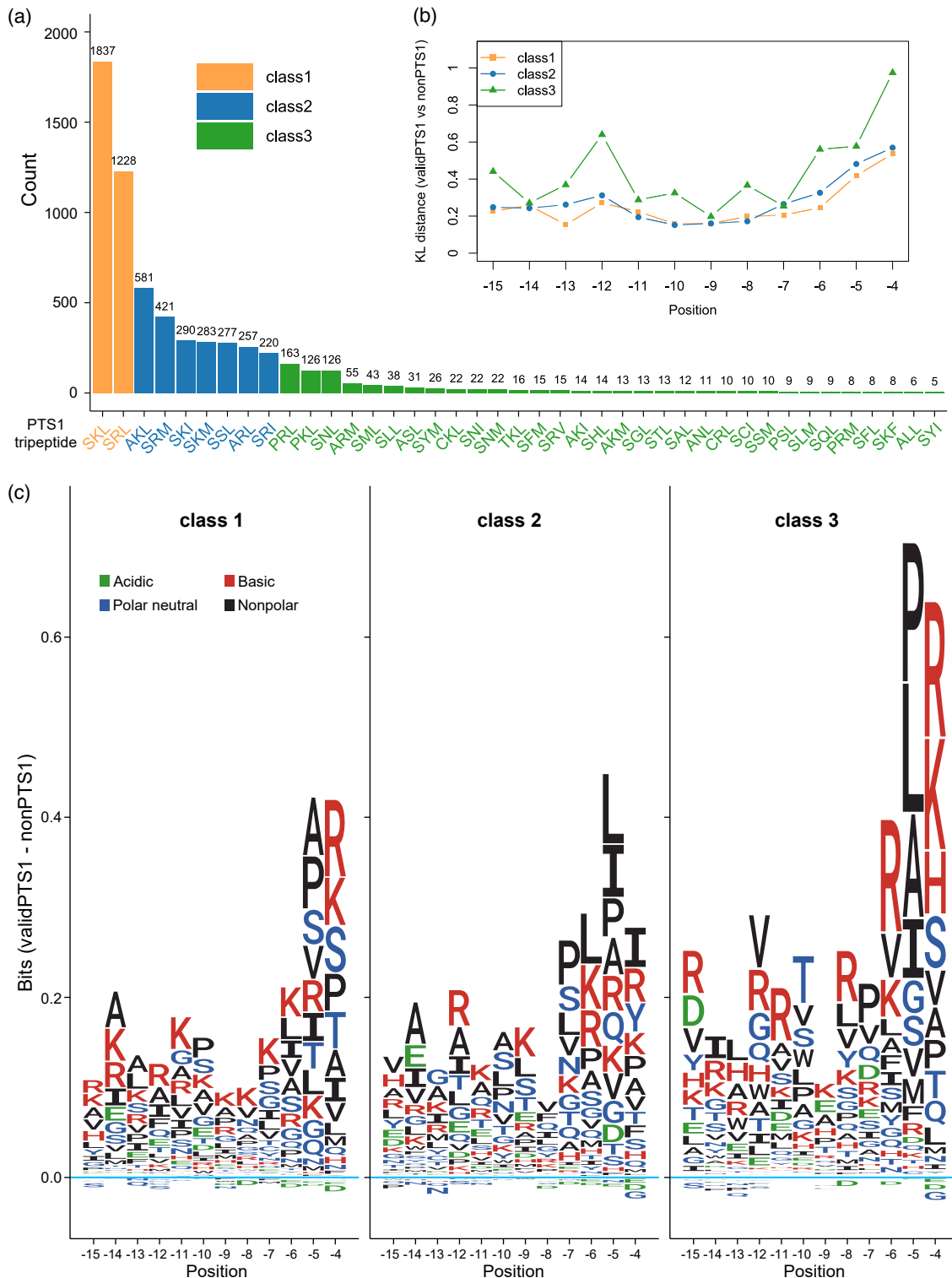
Figure 3. Identification of the upstream enhancing pattern.

(a) Frequencies of different PTS1 tripeptides in the validPTS1 data set (Table S4). The tripeptides were grouped into three classes based on the frequencies of their appearance in the data set.

(b) KL distance analysis of the three classes of the validPTS1 data set using individual amino acids. KL distance was calculated for every upstream position (-15 to -4) between each class of the validPTS1 data set and the nonPTS1 data set.

(c) Seqlogo analysis of the three classes of the PTS1 data set using individual amino acids. The Bits value, calculated by subtracting the value of the nonPTS1 data set from that of each class of the validPTS1 data set, indicates the information content based on amino acid difference at each upstream position (-15 to -4).

upstream elements consistently increased from class 1 to class 3, which is most obvious in the three positions immediately upstream of the tripeptide, −6 to −4 (Figure 3c).

These results suggested that class 3 contains stronger upstream enhancing elements for peroxisome targeting and that the three immediate upstream positions −6 to −4 may play a more determinant role.

Identification and validation of the upstream enhancing sequence pattern

Because class 3 contained relatively rare-occurring PTS1 tripeptides that may have a stronger dependence on upstream enhancing elements for peroxisome targeting, we deduced a potential upstream enhancing pattern, RILVRTKRPRPR, from the most enriched residues at each position of class 3 (Figure 3c).

To choose a weak PTS1 peptide for verification of this enhancing pattern, we analyzed the peroxisome targeting ability of four natural PTS1 peptides ending with weak and non-canonical PTS1 tripeptides: PSL>, SCI>, ALL>, and SYI>, respectively. EMIGRWKRSLAQPSL> is from *Cucumis sativus* HOL3, DWARRFIFGRDSSCI> is from *Dioscorea rotundata* NDB1, RAHVQTEGDRIRALL> is from *Zea mays* HAOX1, and DGAQSLTRPRMRSYI> is from *Betula platyphylla* SDRc. Construct containing mVenus fused to the N-terminus of each peptide (mVenus-PTS1-15aa) was co-expressed with the peroxisomal marker, moxCerulean3-PTS1, in tobacco leaf cells. Confocal microscopy was performed at 36, 48, and 72 h after infiltration to evaluate peroxisome targeting efficiency.

Two of these peptides did not show complete peroxisome targeting at the early time points. EMIGRWKR-SLAQPSL> displayed partial peroxisome targeting at 36 and 48 h and complete targeting at 72 h, and DWARR-FIFGRDSSCI> exhibited no peroxisome targeting at 36 h and partial peroxisome targeting at 48 and 72 h (Figure 4). As discussed below, differences in amino acid properties were discovered at positions −6 to −4 between the two peptides with insufficient peroxisome targeting and the two with complete targeting.

After the enhancing pattern RILVRTKRPRPR was added to the N-terminus of these four PTS1 tripeptides, all four mVenus fusion peptides showed complete peroxisome targeting at all the time points (Figure 4). Thus, our data supported the conclusion that RILVRTKRPRPR serves as an upstream enhancing sequence pattern capable of promoting the efficiency of peroxisome targeting for weak PTS1 tripeptides.

Identification and validation of the upstream inhibiting sequence pattern

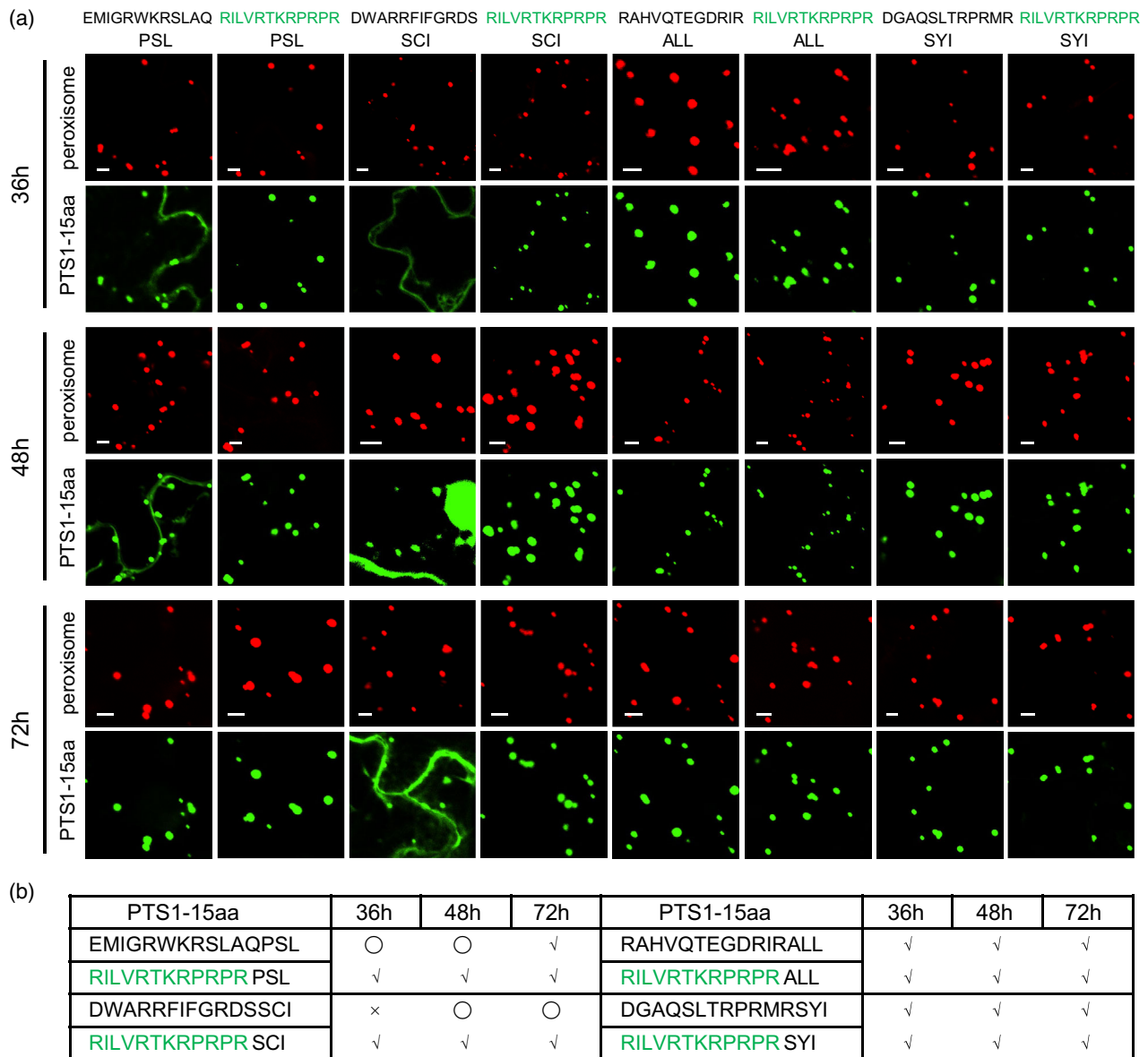The identification of the upstream enhancing pattern prompted us to uncover the upstream inhibiting pattern. We reasoned that, in the nonPTS1 data set, some 15-aa peptides that possess an established PTS1 tripeptide but do not function as PTS1 (Table S5) may contain an upstream inhibiting pattern. To this end, we retrieved 308 such 15-aa peptides and generated an invalidPTS1 data set (Figure 1 and Table S6). Then, we calculated the KL distance between the invalidPTS1 and validPTS1 data sets (Figure 5a). As in the upstream enhancing pattern, positions −6 to −4 in these PTS1-containing non-peroxisomal proteins had relatively higher KL values and may play a determinant role in inhibiting peroxisome targeting. Using the validPTS1 data set as the background for comparison, we also generated a seqlogo of the invalidPTS1 data set, in which acidic and polar neutral residues were found to be enriched (Figure 5b). Consistent with the KL distance analysis, seqlogo analysis showed that the three or four positions immediately upstream of the PTS1 tripeptide contained more information (Figure 5b).

An upstream inhibiting pattern, SSNSDNGLSSFP, was deduced from the most enriched residues at each position of the invalidPTS1 peptides (Figure 5b). This pattern was then tested for its inhibitory effect on peroxisome targeting of the same four weak and non-canonical PTS1 tripeptides used earlier in the study. In stark contrast to the results obtained from the 15-aa peptides containing these four PTS1 tripeptides with their respective natural upstream sequences (Figure 4), the four new 15-aa peptides showed slowed or no peroxisome targeting (Figure 5c,d). With SSNSDNGLSSFP added to its N-terminus, PSL> changed from partial peroxisome targeting at 36 h and full targeting at 72 h to non-peroxisomal targeting at any of the time points (Figure 5c,d). Similarly, SCI> changed from partial to non-peroxisome targeting at 48 h, ALL> changed from full to non-targeting at 36 h, and SYI> changed from full targeting at all time points to non-targeting even at 72 h (Figure 5c,d). These results confirmed the function of SSNSDNGLSSFP as an upstream inhibiting pattern that impedes peroxisome targeting of weak PTS1 tripeptides.

However, this deduced upstream inhibiting sequence pattern does not appear to be very strong. In the seqlogo of the invalid PTS1 peptides, the information content values were only approximately 0.2 even at positions −6 to −4 (Figure 5b) as opposed to approximately 0.6 as found for the upstream enhancing pattern (Figure 3c). When we added the full inhibiting pattern SSNSDNGLSSFP in front of the strong PTS1 tripeptide SKL>, the fusion peptide still showed full peroxisome targeting at all the tested time points (Figure S3). Therefore, SSNSDNGLSSFP is incapable of inhibiting the peroxisome targeting of strong PTS1 tripeptides.

Effects of the polarity and charge of the upstream residues on peroxisome targeting

Our seqlogo analyses using individual residues revealed relative enrichments of basic residues with positive

(a)

| EMIGRWKRSLAQ RILVRTKRPRPR | DWARRFIFGRDS RILVRTKRPRPR | RAHVQTEGDRIR RILVRTKRPRPR | DGAQSLTRPRMR RILVRTKRPRPR |

Figure 4. Validation of the upstream enhancing pattern *in vivo*.

(a) Confocal images were taken from tobacco leaf cells transiently co-expressing the peroxisome marker moxCerulean3-PTS1 and mVenus-PTS1-15aa peptide fusions. Effects of the upstream sequences on the four weak PTS1 tripeptides in peroxisome targeting were tested at three time points. Upstream enhancing pattern is labeled in green. Scale bars = 5 μm.

(b) Comparisons of the peroxisome targeting capabilities of the PTS1-15aa peptides at three time points after tobacco infiltration, based on data presented in (a). The upstream enhancing pattern is labeled in green. $\bigcirc$, ×, and √ indicate partial, no, and complete peroxisome targeting, respectively.

(b)

PTS1-15aa	36h	48h	72h	PTS1-15aa	36h	48h	72h
EMIGRWKRSLAQPSL	$\bigcirc$	$\bigcirc$	√	RAHVQTEGDRIRALL	√	√	√
RILVRTKRPRPR PSL	√	√	√	RILVRTKRPRPR ALL	√	√	√
DWARRFIFGRDSSCI	×	$\bigcirc$	$\bigcirc$	DGAQSLTRPRMRSYI	√	√	√
RILVRTKRPRPR SCI	√	√	√	RILVRTKRPRPR SYI	√	√	√

charges and nonpolar residues throughout the upstream region of the validPTS1 data set (Figure 3c), indicating the importance of residue polarity and charge in peroxisome targeting. To further verify this observation, we grouped the amino acid residues according to their polarity and charge and re-generated KL distance and seqlogos. Both of the new analyses showed similar trends with the analyses using ungrouped individual residues (Figure 6). As a comparison, we also generated seqlogos after grouping residues according to their chemical structure, which resulted in highly noised patterns that showed inconsistent trends among the three classes and were significantly diverged from those obtained from individual residues (Figure S4). These observations suggested that polarity and charge are key characteristics of the upstream enhancing patterns in modulating peroxisome targeting.

The seqlogos based on polarity and charge of the amino acids (Figure 6b) had less noise than the seqlogos based on individual residues (Figure 3c), which was mostly a result of the strong reduction in information content at distant
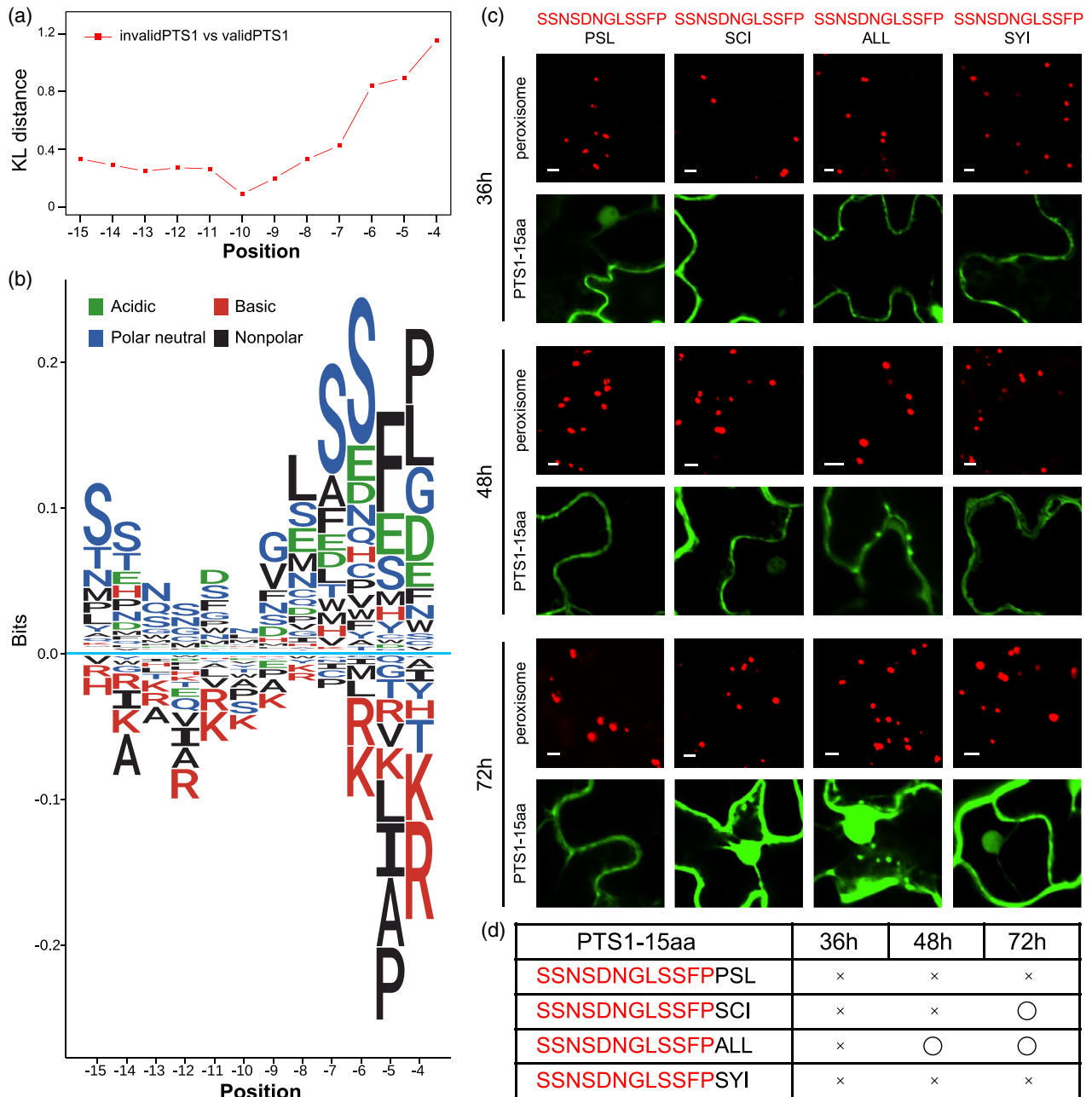
Figure 5. Identification and validation of the upstream inhibiting pattern.
(a) KL distance analysis of the invalidPTS1 data set using individual amino acids. KL distance was calculated for every upstream position (−15 to −4) between the invalidPTS1 and the validPTS1 data sets.
(b) Seqlogo analysis of the invalidPTS1 data set using individual amino acids. The Bits value, calculated by subtracting the value of the validPTS1 data set from that of the invalidPTS1 data set, indicates the information content based on amino acid differences at each upstream position (−15 to −4).
(c) Confocal images of tobacco leaf cells transiently co-expressing the peroxisome marker moxCerulean3-PTS1 and mVenus-PTS1-15aa peptide fusions. Effects of the upstream sequences on the four weak PTS1 tripeptides in peroxisome targeting were tested at three time points after tobacco infiltration. Scale bars = 5 μm.
(d) Comparisons of the peroxisome targeting capabilities of the PTS1-15aa peptides at three time points after tobacco infiltration, based on data presented in (c). The upstream inhibiting pattern is labeled in red. ○, ×, and √ indicate partial, no, and complete peroxisome targeting, respectively.

positions that made positions −6 to −4 more significant. In positions −6 to −4 in class 3, we were able to deduce a 'basic-nonpolar-basic' pattern, among which a nonpolar residue at position −5 had very high information content,

indicating the strong enhancing effect of a nonpolar residue in this position on weak PTS1 tripeptide (Figure 6b).

To experimentally analyze the effect of the polarity and charge of the upstream residues on peroxisome targeting,
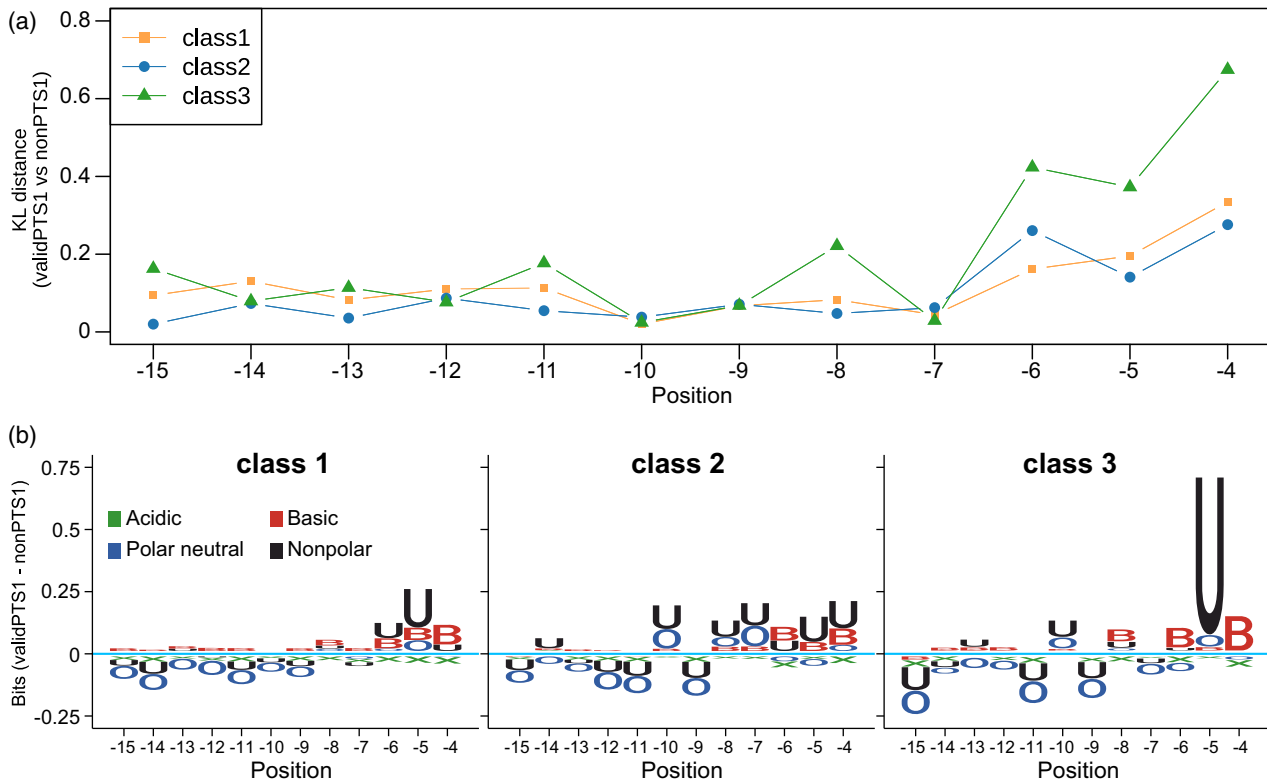
Figure 6. Statistical analysis of the importance of residue polarity and charge in the upstream enhancing pattern.
(a) KL distance analysis of the three classes of the validPTS1 data set based on the polarity and charge of the amino acids. KL distance was calculated for every upstream position (−15 to −4) between each class of the validPTS1 data set and the nonPTS1 data set.
(b) Seqlogo analysis of the three classes of the PTS1 data set based on the polarity and charge of the amino acids. The Bits value, calculated by subtracting the value of the nonPTS1 data set from that of each class of the validPTS1 data set, indicates the information content based on differences in amino acid groups at each upstream position (−15 to −4). Residues with different properties are as follows: basic, K, R, and H; acidic, D and E; nonpolar, A, V, L, I, P, F, W, and M; polar neutral, G, S, T, C, Y, N, and Q.

we deduced an all-basic residue peptide RRHRRKKRRRRR and an all-nonpolar residue peptide VILVAVALPVPV from the enriched residues of class 3 of the validPTS1 data set (Figure 3c). Similarly, an all-acidic peptide EEDDDEDEEEED and an all-polar neutral peptide SSNSSNGSSSSG were deduced from the enriched residues of the invalidPTS1 data set (Figure 5b). These four artificial peptides were respectively fused to the four weak PTS1 tripeptides, PSL>, ALL>, SYI>, and PKL>, and a strong PTS1 tripeptide, SKL> (Figure 7a,b and Figures S5–S7).

Surprisingly, inconsistent with the previous understanding, the all-basic peptide RRHRRKKRRRRR did not show

the anticipated enhancing effect on peroxisome targeting. At 36 h, none of the fusion peptides, even the one that contained SKL>, showed complete peroxisome targeting (Figure 7a–c and Figures S5–S7). By contrast, the all-nonpolar peptide VILVAVALPVPV showed stronger effect in enhancing peroxisome targeting than the all-basic peptide (Figure 7a–c and Figures S5–S7). The all-acidic peptide EEDDDEDEEEED fully inhibited peroxisome targeting of all the tripeptides, including SKL> (Figure 7a–c and Figures S5–S7), suggesting that acidic residues in the upstream region have strong inhibiting effects. The all-polar neutral peptide SSNSSNGSSSSG fully inhibited the

Figure 7. Impact of the polarity, charge, and position of the upstream residues on peroxisome targeting.
(a, b) Confocal images of tobacco leaf cells transiently co-expressing the peroxisome marker moxCerulean3-PTS1 and mVenus-PTS1-15aa peptide fusions. Impacts of the upstream patterns on PSL> and SYI> were tested. Scale bars = 5 μm.
(c) Summary of the peroxisome targeting capabilities of the PTS1-15aa peptides at three time points after tobacco infiltration, based on results presented in (a) and (b) and Figures S5–S7. ○, ×, and √ indicate partial, no, and complete peroxisome targeting, respectively. Sequences in black, red, purple, and green background colors indicate nonpolar, basic, polar neutral, and acidic upstream residues, respectively. (d) Impacts of the upstream patterns on SYI>. Confocal images show tobacco leaf cells transiently co-expressing the peroxisome marker moxCerulean3-PTS1 and mVenus-PTS1-15aa peptide fusions. Scale bars = 5 μm.
(e) Summary of the peroxisome targeting capabilities of the PTS1-15aa peptides at three time points after tobacco infiltration, based on data presented in (d) and Figure S8. ○, ×, and √ indicate partial, no, and complete peroxisome targeting, respectively. Red and green residues are from the deduced upstream inhibiting and enhancing patterns, respectively.
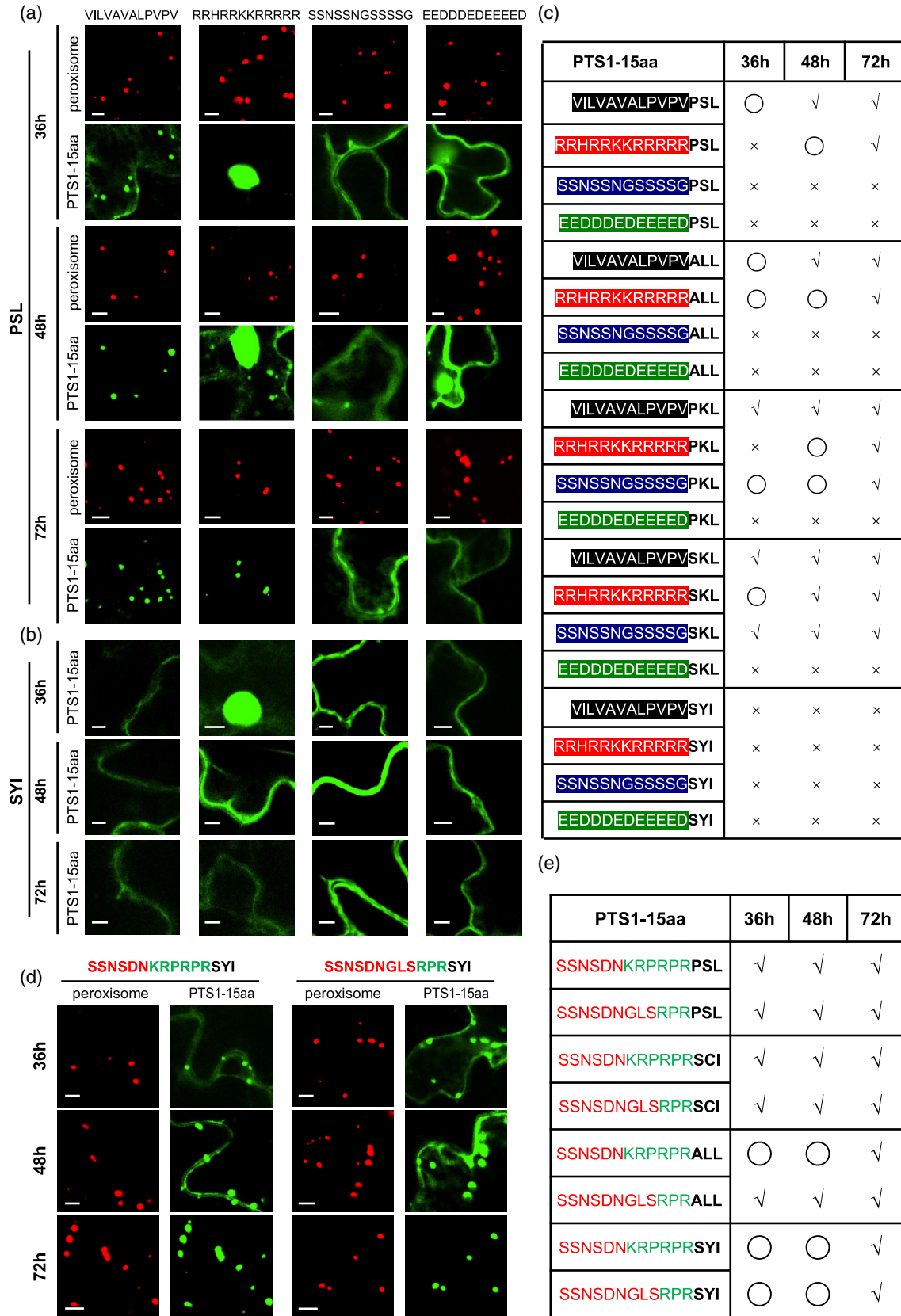
peroxisome targeting of the weak tripeptides, SYI>, PSL> and ALL>, but not on SKL> and PKL> (Figure 7a–c and Figures S5–S7), suggesting that the polar neutral residues in the upstream region may only have moderate inhibiting effects.

Taken together, our data indicated that polarity and charge are key characteristics of the upstream enhancing and inhibiting patterns. When present in all the upstream positions, nonpolar residues have the best enhancing effects on peroxisome targeting, whereas basic residues only have moderate enhancing effects. With respect to the inhibiting effects, acidic residues are the strongest and polar neutral residues have moderate effects.

Determining the core upstream positions within a strong enhancing pattern

Our statistical analyses performed so far suggested that, for both upstream enhancing and inhibiting patterns, positions adjacent to the PTS1 tripeptide have stronger effects than the more distant positions, with the three immediately upstream positions (−6 to −4) having the strongest impact (Figures 3b,c and 5a,b). To further test this, we replaced residues −9 to −4 or −6 to −4 of the inhibiting pattern with those of the enhancing pattern to generate two hybrid peptides, SSNSDN-KRPRPR and SSNSDNGLS-RPR, which were then fused to the four weak PTS1 tripeptides, SYI>, PSL>, ALL>, and SCI> (Figure 7d,e and Figure S8). Both replacements strongly enhanced the peroxisome targeting of the full inhibiting pattern for all the four tripeptides (Figure 7d,e and Figure S8). In comparison with the full enhancing pattern, SSNSDN-KRPRPR achieved similar enhancing effect when added to PSL>, ALL>, and SCI> and mildly reduced the peroxisome targeting efficiency of SYI> and ALL>, whereas SSNSDNGLS-RPR only mildly reduced that of SYI> (Figure 7d,e and Figure S8). We concluded that the three positions immediately upstream of the PTS1 tripeptide are the core upstream positions with the strongest and, at least in some cases, decisive enhancing effect on the targeting ability of the PTS1 tripeptide.

Next, we focused on the core upstream positions to define a strong enhancing pattern. Our seqlogo analysis based on the polarity and charge of the amino acid residues revealed a 'basic-nonpolar-basic' pattern at positions −6 to −4, among which the nonpolar residue had very high information content (Figure 6b), indicating the strong enhancing effect of a nonpolar residue at −5 on weak PTS1 tripeptides. Consistent with this *in silico* observation, the all-basic peptide RRHRRKKRRRRR had a surprisingly weak enhancing effect (Figure 7c), which was obviously weaker than the deduced enhancing pattern RILVRTKRPRPR (Figure 5d), the hybrid peptide SSNSDNGLS-RPR (Figure 7e) containing the inhibiting pattern in all the positions except the core positions, and even the all-nonpolar peptide VIL-VAVALPVPV (Figure 7c). This indicated that, to achieve

strong enhancing effect, positions −6 to −4 should not be all positively charged and it is also pivotal to have a nonpolar residue in position −5. The all-nonpolar peptide VIL-VAVALPVPV (Figure 7c) had weaker enhancing effect than the SSNSDNGLS-RPR peptide that contained the inhibiting pattern in all but the core positions (Figure 7e), demonstrating the significance of basic residues in −6 and −4.

Among the tripeptides used for assessing upstream patterns *in vivo*, SYI> was the only one fully inhibited by the all-basic RRHRRKKRRRRR and the all-nonpolar VILVA-VALPVPV peptides (Figure 7b,c). However, it led to efficient peroxisome targeting when DGAQSLTRPRMR (Figure 5d), RILVRTKRPRPR (Figure 5d), or SSNSDNGLS-RPR (Figure 7d,e) was attached at its N-terminus. Therefore, the nonpolar residue at position −5 and the basic residues at positions −6 and −4 are both indispensable for the peroxisome targeting of SYI>, underscoring the strong and universal enhancing effect of the 'basic-nonpolar-basic' pattern. Consistent with this, RAHVQTEGDRIRALL> and DGAQSLTRPRMRSYI>, the two natural PTS1 peptides with weak tripeptides yet showing strong peroxisome targeting even at the early expression time points, both fit the 'basic-nonpolar-basic' pattern (Figure 4).

In summary, our results provided strong evidence that the three positions immediately upstream of the PTS1 tripeptide are the core upstream positions with the strongest and sometimes decisive effect on peroxisome targeting. The 'basic-nonpolar-basic' pattern at these positions is a strong and universal enhancing pattern.

DISCUSSION

In-depth analysis of the plant PTS1 domain

The peroxisome is amazingly versatile in its metabolic function, as reflected by the existence of many taxa-specific proteins and pathways (Charles et al., 2020; Pan et al., 2020; Parsons, 2004). This versatility makes prediction of peroxisomal proteins highly valuable because it may lead to the discovery of new metabolic pathways in different species (Pan et al., 2020; Reumann & Chowdhary, 2018). PPero and PredPlantPTS1 are two prediction algorithms for plant PTS1-containing proteins (Lingner et al., 2011; Reumann et al., 2012; Wang et al., 2017). However, the capability of these algorithms is still hampered by the incomplete understanding of PTS1 on both the tripeptide and the upstream sequence.

In the present study, we assembled large data sets of PTS1-containing peroxisomal proteins and non-peroxisomal proteins, which enabled us to not only discover 12 low-frequency, non-canonical PTS1 tripeptides, but also retrieve hundreds of PTS1 peptides with rare-occurring PTS1 tripeptides and non-peroxisomal peptides with established PTS1 tripeptides (Tables S4 and S6). From these data sets, we deduced upstream enhancing and

inhibiting sequence patterns and tested their impacts on the function of PTS1 tripeptides. Our study provided strong evidence to correct the previous understanding that positive charges constitute the upstream enhancing pattern. Our systematic analyses of the polarity, charge, and position of upstream amino acids led to the identification of core positions (−6 to −4) that constitute a strong and universal enhancing pattern of 'basic-nonpolar-basic' for peroxisome targeting of the PTS1 tripeptides. These findings have significantly expanded our knowledge of the PTS1 domain in plants and likely other eukaryotes as well.

The upstream enhancing and inhibiting patterns mainly impact rare PTS1 tripeptides

The large number of plant PTS1-containing proteins collected in the present study enabled us to categorize them into three classes based on the frequency of appearance of the PTS1 tripeptides. Comparative analysis provided statistical evidence that clearly supported the inverse relationships between the strength of the upstream enhancing elements and the frequency of occurrence of the tripeptides (Figure 3). That is, the less frequent tripeptides are weaker in targeting strength and thus more dependent on the upstream elements.

Our *in vivo* localization experiments also demonstrated that both the upstream enhancing and inhibiting patterns are more influential on the rare and weak tripeptides. This is consistent with the notion that these rare tripeptides are neither strong enough by themselves to efficiently target the protein to peroxisomes, nor capable of overcoming the inhibitory effect of upstream inhibiting elements. It is also possible that the enhancing pattern cannot further improve the targeting of the strong tripeptides, which by themselves are already sufficient for efficient peroxisome targeting.

All class 3 proteins in our peroxisomal PTS1 protein data set contain weak PTS1 tripeptides, which are expected to depend more on the upstream enhancing pattern for peroxisome targeting. Over 20% of them carry the strong enhancing pattern of 'basic-nonpolar-basic' at positions −6 to −4 (Figure S9 and Table S1). Most of the proteins in this category, such as short-chain dehydrogenase/reductase c (SDRc), hydroxy-acid oxidase 1 (HAOX1), 4-coumarate:CoA ligase 2 (4CL2), and small thioesterase 5 (ST5), participate in minor peroxisomal functions. A few others, such as abnormal inflorescence meristem 1 (AIM1), multifunctional protein 2 (MFP2), malate synthase (MLS), and 12-oxophytodienoate reductase 3 (OPR3), are enzymes in core peroxisomal pathways (Table S1).

In the present study, we also found acidic residues at the core upstream positions −6 to −4 to have strong inhibitory effect on peroxisome targeting of class 3 proteins, which may likely result in partial peroxisomal localization for some proteins. None of the class 3 proteins contain two or three acidic residues at these positions, which would

otherwise strongly inhibit peroxisome targeting. Only approximately 5.4% of class 3 proteins contain one acidic residue at these positions (Figure S9 and Table S1) and, with one exception (*Antirrhinum majus* AIM1), all of them participate in minor peroxisomal functions. Examples include NAD(P)H dehydrogenase B1 (NDB1), mitochondrial intermembrane space assembly machinery 40 (MIA40), and acyl-activating enzyme 17 (AAE17); among them, NDB1 and MIA40 have been shown to dually localize to mitochondria and peroxisomes (Carrie et al., 2008, 2010) (Table S1). In addition, three natural PTS1-15aa peptides, DWTRRFIFGRDSSCI (*Digitaria exilis* NDB1), DWMKRFIFGRDSSCI (*Ananas comosus* NDB1), and DWARRFIFGRDSSCI (*Dioscorea rotundata* NDB1), all of which contain an acidic amino acid at the core upstream positions, only exhibit partial peroxisome targeting even after 72 h of expression (Figure 2b).

Limitations in defining strong upstream inhibiting pattern in the present study

The upstream inhibiting sequence pattern SSNSDNGLSSFP deduced in the present study was incapable of inhibiting the peroxisome targeting of the strong PTS1 tripeptide SKL (Figure S3). Thus, it may not be a very strong inhibiting pattern. The nonPTS1 peptides used in this study are from TFs. Although never been demonstrated, we cannot exclude the possibility that some TFs are capable of localizing to peroxisomes. It is also possible that some TFs contain PTS1 peptides but do not target to peroxisomes for reasons other than the presence of upstream inhibiting patterns. For example, there may be strong nuclear targeting signals or transmembrane domains, or the PTS1 peptide is not exposed to the surface because of protein folding or interaction with other proteins. Moreover, despite the large size (308 sequences) of the invalidPTS1 data set, most of the peptides in this data set end with rare non-canonical tripeptides that belong to class 3. Only 73 samples belong to classes 1 and 2, which may contribute to the lack of strength of the deduced inhibiting pattern. To define the strong upstream inhibiting pattern, a larger invalidPTS1 data set that only contains samples with strong PTS1 tripeptides is needed. However, soluble non-peroxisomal proteins with strong PTS1 tripeptide are scarce. Continuous sequencing of various plant genomes may allow us to identify more plant PTS1-containing non-peroxisomal proteins and possibly solve this problem in the future.

Structural information of PEX5 may provide additional clues to PTS1 prediction

Structural studies of human PEX5 and its interaction with PTS1 peptides have found that three structural components of a PTS1 peptide are involved in its binding to PEX5: the terminal carboxyl group, the peptide backbone, and the sidechains (Gatto et al., 2000; Reumann et al., 2016; Stanley et al., 2006). Among them, the sidechains

contribute to sequence specificity by fitting into multiple binding pockets of PEX5 (Gatto et al., 2000; Reumann et al., 2016; Stanley et al., 2006). Furthermore, PEX5 changes its conformation after PTS1 binding (Gatto et al., 2000; Stanley et al., 2006). Structural analysis of human PEX5 revealed that it has a flexible TPR domain for PTS1 binding and can adapt its conformation differentially to different PTS1 cargo with different receptor binding affinities (Fodor et al., 2015). This flexibility may explain the high tolerance of amino acid variability and taxa specificities of the PTS1 domain (Ghosh & Berg, 2010; Reumann et al., 2016; Sampathkumar et al., 2008).

Given that PEX5 provides the structural constraint that may determine the characteristics of PTS1, a precise understanding of how PEX5 recognizes and interacts with PTS1 peptides could lay the foundation for predicting PTS1 peptides through structure-based algorithms that complement the sequence-based algorithms. As a result of the rapid progress made with respect to using artificial intelligence to predict protein structure and interaction, this approach may become feasible in the foreseeable future (Humphreys et al., 2021; Jumper et al., 2021).

Implications for deciphering auxiliary targeting signals for peroxisomes in other eukaryotes and those for other types of organelles

Orthologs of the PTS1 receptor PEX5 from different organisms share high sequence similarities and a largely conserved function (Reumann et al., 2016; Wimmer et al., 1998). The general sequence properties of the PTS domain are shared between fungi, mammals, and plants (Brocard & Hartig, 2006; Lametschwandtner et al., 1998; Reumann & Chowdhary, 2018). Hence, our findings regarding the features of the upstream targeting elements will also shed light on the understanding of the PTS1 domain in other eukaryotic systems. However, PEX5 homologs differ significantly among fungi, mammals, and plants in their affinity for specific tripeptides, which is reflected by experimentally determined kingdom-specific PTS1 consensus motifs (Brocard & Hartig, 2006; Emanuelsson et al., 2003; Lametschwandtner et al., 1998; Reumann & Chowdhary, 2018). Hence, the characteristics of the PTS1 domain may be kingdom specific as well and should be analyzed in a lineage specific manner. Nonetheless, the large-scale strategies employed in the present study could be applied to dissecting auxiliary signals for peroxisome targeting in other eukaryotic systems.

In the present study, the upstream enhancing and inhibiting patterns for PTS1 in peroxisome targeting were deduced from large collections of peroxisomal proteins with low-frequency PTS1 tripeptides and non-peroxisomal proteins with established PTS1 tripeptides. In addition to peroxisomes, other organelles, such as mitochondria, chloroplasts, the secretory pathway, and the nucleus, also utilize protein targeting signals formed by short peptide motifs for which the targeting strengths may be influenced by adjacent sequence. Thus, the principles and strategies employed in the present study may also be applied to the analysis of auxiliary enhancing and inhibiting targeting signals for these organelles.

EXPERIMENTAL PROCEDURES

Homology-based searches for mutual best-match proteins

One hundred seventy-seven species covering all the main clades of angiosperms were selected from species with completely sequenced genomes (https://www.plabipd.de). Peptide sequences were downloaded from the databases: NCBI Genome (https://www.ncbi.nlm.nih.gov/genome), Phytozome v13 (https://phytozome.jgi.doe.gov), and Ensembl Plants release 49 (https://plants.ensembl.org).

A protein was selected only if it was the 'mutual best-match' in the two-way BLAST search between its corresponding plant species and Arabidopsis. Searches for mutual homologs were coducted using DIAMOND v2.0.6 with sensitive model and *e* value 1×10^{-10} (Buchfink et al., 2021) to identify the mutual best-match for each reference protein. Protein sequence identity of 30% was a threshold for a protein to be considered.

KL distance calculation

KL divergence quantifies how much one probability distribution differs from another probability distribution (Kullback & Leibler, 1951). We used two discrete probability distributions, $P(z)$ and $Q(z)$, to represent the distribution of amino acids at each position of the sequence. KL divergence from Q to P is defined as:

$$D_{\text{KL}}\left(P(z)\middle\|Q(z)\right) \equiv \sum_z P(z)\log\left[\frac{P(z)}{Q(z)}\right] \qquad (1)$$

Note that KL divergence is not symmetrical, which means that $D_{\text{KL}}\left(P\middle\|Q\right) \neq D_{\text{KL}}\left(Q\middle\|P\right)$. To compare the distribution of amino acids at each site, we used a symmetric KL divergence defined as:

$$\frac{1}{2}D_{\text{KL}}\left(P\middle\|Q\right) + \frac{1}{2}D_{\text{KL}}\left(Q\middle\|P\right) \qquad (2)$$

Seqlogo analysis

A seqlogo consists of a stack of amino acid letters at each position. The height of the letters indicates the information content at this position.

The information content of position i is calculated as:

$$I_i = \log_2 20 - H_i \qquad (3)$$

H_i is the Shannon entropy of position i. 20 indicates the total types of amino acids; and $f_{b,i}$ is defined as the relative frequency of amino acid b at position i. H_i can be calculated as:

$$H_i = -\sum_{b=1}^{20} f_{b,i} \times \log_2 f_{b,i} \qquad (4)$$

Therefore, the height of amino acid b in column i is calculated as $f_{b,i} \times I_i$.

Gene cloning and plasmid construction

For tobacco transient protein expression, fusions between mVenus and the PTS1 peptides were obtained by overlapping PCR (primers shown in Table S7). Briefly, two PCR reactions were

performed to generate each mVenus-PTS1-15aa fusion: the first reaction used primers F and R1 with mVenus coding sequence as the template, and the second reaction used primers F and R2 with products from the first reaction as the template. The fusion product was then cloned into the pCAMBIA1300-mVenus vector, which already contained the 35S constitutive promoter and cut by *Xba*l and *Sac*l (New England Biolabs, Beijing, China), to replace mVenus, using the ClonExpress II One Step Cloning Kit (Vazyme, Nanjing, China).

To generate the peroxisome marker moxCerulean3-PTS1, a SKL tripeptide was fused to the C-terminus of the moxCerulean3 fluorescent protein before the fusion construct was cloned into the pGWB545 vector backbone (Nakagawa et al., 2007).

Transient protein expression and *in vivo* targeting analysis

The constructs were first transformed into *Agrobacterium tumefaciens* strain GV3101 (pMP90) via heat shock (Rainer & Willmitzer, 1988). Transient protein expression in tobacco (*Nicotiana tabacum*) leaves followed by confocal microscopy to analyze protein targeting was carried out as described previously (Pan et al., 2014). A Fluoview FV3000 confocal laser-scanning microscope (Olympus, Tokyo, Japan) was used for image capturing, where moxCerulean3 was excited with 445-nm lasers and detected at 460–500 nm and mVenus was excited with 514-nm lasers and detected at 530–630 nm.

ACKNOWLEDGEMENTS

AUTHOR CONTRIBUTIONS

HJ, JH, and RP co-conceptualized and co-supervised the study. YF, RP, RX, XS, DA, WL, and RZ prepared the data sets. QD and RX generated the constructs. QD and RX performed peroxisome targeting analysis. HL, RX, and HJ performed the statistical analysis and generated the statistics and figures. JH, RP, HJ, and YF co-wrote the manuscript.

CONFLICT OF INTERESTS

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of the present study are available in the Supporting information.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Table S1. Mutual best-match data set for plant peroxisomal PTS1 proteins.

Table S2. Mutual best-match data set for plant proteins containing both PTS1 and PTS2.

Table S3. Mutual best-match data set for non-peroxisomal proteins.

Table S4. The validPTS1 data set.

Table S5. The nonPTS1 data set.

Table S6. The invalidPTS1 data set.

Table S7. Primers used in the present study.

Figure S1. PTS1-3aa of mutual best-match proteins containing both PTS1 and PTS2.

Figure S2. PTS1 tripeptide composition in different orthologous protein groups.

Figure S3. Targeting analysis of the impact of the deduced upstream inhibiting pattern on SKL>.

Figure S4. Seqlogo analysis of the three classes of the PTS1 data set using amino acid groups of specific chemical structure.

Figure S5. Targeting analysis of the impact of upstream polarity and charge on ALL>.

Figure S6. Targeting analysis of the impact of upstream polarity and charge on PKL>.

Figure S7. Targeting analysis of the impact of upstream polarity and charge on SKL>.

Figure S8. Targeting analysis of the impact of the position of the upstream residues on PSL>, SCI> and ALL>.

Figure S9. Distribution of class 3 proteins with different patterns at core upstream positions.

REFERENCES

Bongcam, V., Petétot, J.M.D.C., Mittendorf, V., Robertson, E.J., Leech, R.M., Qin, Y.M. *et al.* (2000) Importance of sequences adjacent to the terminal tripeptide in the import of a peroxisomal *Candida tropicalis* protein in plant peroxisomes. *Planta*, **211**, 150–157.

Brocard, C. & Hartig, A. (2006) Peroxisome targeting signal 1: is it really a simple tripeptide? *Biochimica et Biophysica Acta - Molecular Cell Research*, **1763**, 1565–1573.

Buchfink, B., Reuter, K. & Drost, H.G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, **18**, 366–368.

Carrie, C., Giraud, E., Duncan, O., Xu, L., Wang, Y., Huang, S. *et al.* (2010) Conserved and novel functions for *Arabidopsis thaliana* MIA40 in assembly of proteins in mitochondria and peroxisomes. *Journal of Biological Chemistry*, **285**, 36138–36148.

Carrie, C., Murcha, M.W., Kuehn, K., Duncan, O., Barthet, M., Smith, P.M. *et al.* (2008) Type II NAD(P)H dehydrogenases are targeted to mitochondria and chloroplasts or peroxisomes in *Arabidopsis thaliana*. *FEBS Letters*, **582**, 3073–3079.

Charles, K.N., Shackelford, J.E., Faust, P.L., Fliesler, S.J., Stangl, H. & Kovacs, W.J. (2020) Functional peroxisomes are essential for efficient cholesterol sensing and synthesis. *Frontiers in Cell and Developmental Biology*, **8**, 1115.

Chowdhary, G., Kataya, A.R.A., Lingner, T. & Reumann, S. (2012) Non-canonical peroxisome targeting signals: identification of novel PTS1 tripeptides and characterization of enhancer elements by computational permutation analysis. *BMC Plant Biology*, **12**, 1–14.

Chu, C.C., Swamy, K. & Li, H.M. (2020) Tissue-specific regulation of plastid protein import via transit-peptide motifs. *Plant Cell*, **32**, 1204–1217.

Corpas, F.J. (2019) Peroxisomes in higher plants: an example of metabolic adaptability. *Botany Letters*, **166**, 298–308.

Distel, B., Gould, S.J., Voorn-Brouwer, T., van der Berg, M., Tabak, H.F. & Subramani, S. (1992) The carboxyl-terminal tripeptide serine-lysine-leucine of firefly luciferase is necessary but not sufficient for peroxisomal import in yeast. *The New Biologist*, **4**, 157–165.

Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols*, **2**, 953–971.

Emanuelsson, O., Elofsson, A., Von Heijne, G. & Cristóbal, S. (2003) In silico prediction of the peroxisomal proteome in fungi, plants and animals. *Journal of Molecular Biology*, **330**, 443–456.

Falter, C., Thu, N.B.A., Pokhrel, S. & Reumann, S. (2019) New guidelines for fluorophore application in peroxisome targeting analyses in transient plant expression systems. *Journal of Integrative Plant Biology*, **61**, 884–899.

Fodor, K., Wolf, J., Reglinski, K., Passon, D.M., Lou, Y., Schliebs, W. et al. (2015) Ligand-induced compaction of the PEX5 receptor-binding cavity impacts protein import efficiency into peroxisomes. *Traffic (Copenhagen, Denmark)*, **16**, 85–98.

Gabaldón, T. (2010) Peroxisome diversity and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 765–773.

Gatto, G.J., Geisbrecht, B.V., Gould, S.J. & Berg, J.M. (2000) Peroxisomal targeting signal-1 recognition by the TPR domains of human PEX5. *Nature Structural Biology*, **7**, 1091–1095.

Ghosh, D. & Berg, J.M. (2010) A proteome-wide perspective on peroxisome targeting signal 1(PTS1)-Pex5p affinities. *Journal of the American Chemical Society*, **132**, 3973–3979.

Honsho, M., Okumoto, K., Tamura, S. & Fujiki, Y. (2020) Peroxisome biogenesis disorders. *Advances in Experimental Medicine and Biology*, **1299**, 45–54.

Hu, J., Baker, A., Bartel, B., Linka, N., Mullen, R.T., Reumann, S. et al. (2012) Plant peroxisomes: biogenesis and function. *Plant Cell*, **24**, 2279–2303.

Humphreys, I.R., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S. et al. (2021) Computed structures of core eukaryotic protein complexes. *Science*, **374**, 1340.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O. et al. (2021) Highly accurate protein structure prediction with alphafold. *Nature*, **596**(7873), 583–589.

Kragler, F., Lametschwandtner, G., Christmann, J., Hartig, A. & Harada, J.J. (1998) Identification and analysis of the plant peroxisomal targeting signal 1 receptor NtPEX5. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 13336–13341.

Kullback, S. & Leibler, R.A. (1951) On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86.

Lametschwandtner, G., Brocard, C., Fransen, M., Van Veldhoven, P., Berger, J. & Hartig, A. (1998) The difference in recognition of terminal tripeptides as peroxisomal targeting signal I between yeast and human is due to different affinities of their receptor Pex5p to the cognate signal and to residues adjacent to it. *Journal of Biological Chemistry*, **273**, 33635–33643.

Lingner, T., Kataya, A.R., Antonicelli, G.E., Benichou, A., Nilssen, K., Chen, X.Y. et al. (2011) Identification of novel plant peroxisomal targeting signals by a combination of machine learning methods and in vivo subcellular targeting analyses. *Plant Cell*, **23**, 1556–1572.

Ma, C. & Reumann, S. (2008) Improved prediction of peroxisomal PTS1 proteins from genome sequences based on experimental subcellular targeting analyses as exemplified for protein kinases from Arabidopsis. *Journal of Experimental Botany*, **59**, 3767–3779.

Mullen, R.T., Lee, M.S., Flynn, C.R. & Trelease, R.N. (1997) Diverse amino acid residues function within the type 1 peroxisomal targeting signal: implications for the role of accessory residues upstream of the type 1 peroxisomal targeting signal. *Plant Physiology*, **115**, 881–889.

Murcha, M.W., Kmiec, B., Kubiszewski-Jakubiak, S., Teixeira, P.F., Glaser, E. & Whelan, J. (2014) Protein import into plant mitochondria: signals, machinery, processing, and regulation. *Journal of Experimental Botany*, **65**, 6301–6335.

Nakagawa, T., Suzuki, T., Murata, S., Nakamura, S., Hino, T., Maeo, K. et al. (2007) Improved gateway binary vectors: high-performance vectors for creation of fusion constructs in transgenic analysis of plants. *Bioscience, Biotechnology and Biochemistry*, **71**, 2095–2100.

Neuberger, G., Maurer-Stroh, S., Eisenhaber, B., Hartig, A. & Eisenhaber, F. (2003) Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences. *Journal of Molecular Biology*, **328**, 567–579.

Pan, R. & Hu, J. (2018) Proteome of plant peroxisomes. *Subcellular Biochemistry*, **89**, 3–45.

Pan, R., Kaur, N. & Hu, J. (2014) The Arabidopsis mitochondrial membrane-bound ubiquitin protease UBP27 contributes to mitochondrial morphogenesis. *Plant Journal*, **78**, 1047–1059.

Pan, R., Liu, J., Wang, S. & Hu, J. (2020) Peroxisomes: versatile organelles with diverse roles in plants. *New Phytologist*, **225**, 1410–1427.

Pan, R., Reumann, S., Lisik, P., Tietz, S., Olsen, L.J. & Hu, J. (2018) Proteome analysis of peroxisomes from dark-treated senescent Arabidopsis leaves. *Journal of Integrative Plant Biology*, **60**, 1028–1050.

Parsons, M. (2004) Glycosomes: parasites and the divergence of peroxisomal purpose. *Molecular Microbiology*, **53**, 717–724.

Rainer, H. & Willmitzer, L. (1988) Storage of competent cells for *Agrobacterium* transformation. *Nucleic Acids Research*, **16**, 9877.

Ramirez, R.A., Espinoza, B. & Kwok, E.Y. (2014) Identification of two novel type 1 peroxisomal targeting signals in *Arabidopsis thaliana*. *Acta Histochemica*, **116**, 1307–1312.

Reumann, S. (2004) Specification of the peroxisome targeting signals type 1 and type 2 of plant peroxisomes by bioinformatics analyses. *Plant Physiology*, **135**, 783–800.

Reumann, S. & Bartel, B. (2016) Plant peroxisomes: recent discoveries in functional complexity, organelle homeostasis, and morphological dynamics. *Current Opinion in Plant Biology*, **34**, 17–26.

Reumann, S., Buchwald, D. & Lingner, T. (2012) PredPlantPTS1: a web server for the prediction of plant peroxisomal proteins. *Frontiers in Plant Science*, **3**, 1–10.

Reumann, S. & Chowdhary, G. (2018) Prediction of peroxisomal matrix proteins in plants. *Subcellular Biochemistry*, **89**, 125–138.

Reumann, S., Chowdhary, G. & Lingner, T. (2016) Characterization, prediction and evolution of plant peroxisomal targeting signals type 1 (PTS1s). *Biochimica et Biophysica Acta - Molecular Cell Research*, **1863**, 790–803.

Sampathkumar, P., Roach, C., Michels, P.A.M. & Hol, W.G.J. (2008) Structural insights into the recognition of peroxisomal targeting signal 1 by *Trypanosoma brucei* peroxin 5. *Journal of Molecular Biology*, **381**, 867–880.

Stanley, W.A., Filipp, F.V., Kursula, P., Schüller, N., Erdmann, R., Schliebs, W. et al. (2006) Recognition of a functional peroxisome type 1 target by the dynamic import receptor Pex5p. *Molecular Cell*, **24**, 653–663.

Teufel, F., Almagro Armenteros, J.J., Johansen, A.R., Gíslason, M.H., Pihl, S.I., Tsirigos, K.D. et al. (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*. https://doi.org/10.1038/s41587-021-01156-3

Wagih, O. (2017) Ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, **33**, 3645–3647.

Wang, J., Wang, Y., Gao, C., Jiang, L. & Guo, D. (2017) PPero, a computational model for plant PTS1 type peroxisomal protein prediction. *PLoS One*, **12**, e0168912.

Wimmer, C., Schmid, M., Veenhuis, M. & Gietl, C. (1998) The plant pts1 receptor: similarities and differences to its human and yeast counterparts. *Plant Journal*, **16**, 453–464.