Running Head:
The Golgi proteome and novel putative glycosyltransferases

Corresponding Author:
Paul Dupree
Department of Biochemistry, University of Cambridge
Tennis Court Road
Cambridge CB2 1QW, United Kingdom
Phone: +44 (0)1223 333340
Fax: +44 (0)1223 333345
E-mail: pd101@cam.ac.uk

Research Area:
Systems biology, molecular biology, and gene regulation.
Putative glycosyltransferases and other plant Golgi apparatus proteins are revealed by LOPIT proteomics

Nino Nikolovski¹,³, Denis Rubtsov¹, Marcelo P Segura¹, Godfrey P Miles¹,⁴, Tim J. Stevens¹, Tom PJ Dunkley¹,³, Sean Munro², Kathryn S Lilley¹,³, Paul Dupree¹,*

¹ Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, United Kingdom; ² MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK; ³ Cambridge Centre for Proteomics, Cambridge Systems Biology Centre, Department of Biochemistry, University of Cambridge, Cambridge, CB2 1QR, UK
Footnotes:
The work presented in this paper was supported by grants from the BBSRC: BB/G016240/1 BBSRC Sustainable Energy Centre Cell Wall Sugars Programme (BSBEC) and BBS/B/10684 to PD and KSL, and from the National Commission of Scientific and Technological Investigation of Chile (CONICYT) to MPS, as well as funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under the Grant Agreement 211982 (RENEWALL).

The authors responsible for distribution of materials integral to the findings presented in this article in accordance with the Journal policy described in the Instructions for Authors (http://www.plantphysiol.org) are: N. Nikolovski (nn254@cam.ac.uk) and P. Dupree (pd101@cam.ac.uk).

‡Current address: U.S. Department of Agriculture; Agricultural Research Service; Wapato, WA USA

*Corresponding author. e-mail: pd101@cam.ac.uk

Abbreviations:
PM, Plasma Membrane; ER, Endoplasmic Reticulum; TGN, Trans-Golgi Network; GT, GlycosylTransferase; SAM, S-AdenosylMethionine; MT, MethylTransferase; TMD, TransMembrane Domain; LOPIT, Localization of Organelle Proteins by Isotope Tagging; iTRAQ, isobaric Tag for Relative and Absolute Quantitation; PLS, Partial Least Squares; HMM, Hidden Markov Model; CAZy, Carbohydrate-Active enZYmes; GES, Goldman, Engelman and Steitz.
ABSTRACT

The Golgi apparatus is the central organelle in the secretory pathway and plays key roles in glycosylation, protein sorting and secretion in plants. Enzymes involved in the biosynthesis of complex polysaccharides, glycoproteins and glycolipids are located in this organelle, but the majority of them remain uncharacterized. Here, we studied the Arabidopsis membrane proteome with a focus on the Golgi apparatus using localization of organelle proteins by isotope tagging (LOPIT). By applying multivariate data analysis to a combined dataset of two new and two previously published LOPIT experiments, we identify the subcellular localization of 1110 proteins with a high confidence. These include 197 Golgi apparatus proteins, 79 of which have not been previously localized by a high-confidence method, as well as the localization of 304 endoplasmic reticulum and 208 plasma membrane proteins. Comparison of the hydrophobic domains of the localized proteins showed that the single-span transmembrane domains have unique properties in each organelle. Many of the novel Golgi-localized proteins belong to uncharacterized protein families. Structure-based homology analysis identified twelve putative Golgi glycosyltransferase (GT) families that have no functionally characterized members, and are therefore not yet assigned to a CAZy GT family. The substantial numbers of these putative GTs leads us to estimate that the true number of plant Golgi GTs might be a third above those currently annotated. Other newly identified proteins are likely to be involved in transport and inter-conversion of nucleotide sugar substrates, as well as polysaccharide and protein modification.

KEY WORDS: Golgi apparatus; organelle proteomics; glycosyltransferases; plant cell wall biosynthesis
INTRODUCTION

The Golgi apparatus is the central organelle in the secretory pathway, and in higher plants is involved in the biosynthesis and transport of cell wall matrix polysaccharides, glycoproteins, proteoglycans and glycolipids, as well as in protein trafficking to different subcellular compartments. The last decade has produced substantial findings on the function of the Golgi apparatus: insights into the protein trafficking at the ER/Golgi interface, the Golgi structural maintenance, its involvement in endocytosis and its behavior during cell division (for a recent review see Faso et al., 2009). However, despite its importance, only a small proportion of the Golgi proteome has been studied: relatively few Golgi proteins have been localized and even fewer have been functionally characterized.

The Golgi apparatus is thought to contain a large and diverse group of membrane bound glycosyltransferases (GTs). The current view is that different GT activities are required for synthesis of the linkage between different donor and acceptor sugars. Having in mind the diversity of linkage types found in cell wall polysaccharides, the number of different GTs involved is likely to be very large. For instance, it has been estimated that for the biosynthesis of pectin alone, the action of 65 different enzymatic activities is needed (Caffall and Mohnen, 2009). By the end of the year 2011, 468 Arabidopsis sequences had been annotated in the Carbohydrate-Active enZYmes (CAZy) GT database (Cantarel et al., 2009; http://www.cazy.org). We estimate that two thirds of these CAZy classified GTs may be targeted to the Golgi. The remaining third are cytosolic or plastidic enzymes involved in processes including, secondary metabolism or starch synthesis. The reported sequences are classified into 43 CAZy families based on amino acid sequence similarities within which at least one member has been biochemically characterized. Each family is likely to have a common structural fold, and three-dimensional (3-D) structures have been resolved for 20 of these 43 families. These are divided mostly into two structural classes, having either a GT-A fold or a GT-B fold (Unligil and Rini, 2000; Bourne and Henrissat, 2001). Moreover, most of the structurally uncharacterized GT families are predicted to adopt either the GT-A or GT-B fold based on 3-D structural homology modeling (Coutinho et al., 2003; Lairson et al., 2008). Despite this conserved 3-D structure, different GT families have very low or undetectable sequence similarities. Consequently, predicting novel GTs based solely on their amino acid sequence similarities is not always achievable, and structural homology searches have also proven useful (Hansen et al., 2009).
The length and properties of the Trans-Membrane Domain (TMD) of endomembrane proteins appear to play a role in protein sorting and location within the secretory pathway and can be used to predict protein localization (Hanton et al., 2005; Sharpe et al., 2010). In order to perform such predictions, a high number of experimentally localized proteins is required, but only limited data sets have been available for plants to date.

In order to identify the most abundant CAZy classified GTs as well as novel putative GTs, in this work we rigorously extended our proteomic studies of the Golgi apparatus. We have previously developed a high throughput MS-based quantitative proteomics technique for subcellular localization of membrane proteins (LOPIT, Dunkley et al., 2004; Dunkley et al., 2006). Here we report new LOPIT datasets and apply a new method of combining them with published LOPIT datasets, localizing an unprecedented number of plant organelle proteins. We have analyzed the TMD properties of the proteins assigned to the ER, Golgi and PM, and determined the organelle-specific features. Structural prediction analysis of the Golgi-localized proteins with unknown function assessed the protein sequences for the potential to fold similarly to known GT structures. We found that the Golgi contains a substantial number of candidate GT families that have no characterized functions. These results yield a broader understanding of the Golgi function and its biochemical properties.
RESULTS

Integration of LOPIT datasets increases coverage of the Golgi proteome

We have previously described the LOPIT approach for high-throughput determination of the steady state localization of membrane proteins (Dunkley et al., 2004; Dunkley et al., 2006). In this quantitative proteomic technique, proteins can be confidently assigned to organelles according to the pattern of their distribution on density gradients (Dunkley et al., 2006). The main aim of the work described here was to increase the number of identified Golgi-localized proteins and to find proteins putatively involved in the biosynthesis of the plant cell wall. Earlier additional replicate experiments identified relatively few novel proteins (Sadowski et al., 2008). Moreover, the additional fractionation data from replicate experiments is not easily combined to strengthen localization predictions because there are often incomplete observations of proteins in each dataset.

Many of the proteins involved in sugar nucleotide metabolism and membrane trafficking are peripheral. In the LOPIT experiments by Dunkley et al., (2006), these were removed by the carbonate washing of the isolated membranes (LOPIT experiments 1 and 2, Fig. 1). To increase the likelihood of identifying novel proteins, we carried out two additional LOPIT experiments with samples enriched in peripheral and luminal proteins (LOPIT experiments 3 and 4, Fig. 1). Arabidopsis liquid-grown callus membrane fractions were prepared using conditions to preserve the association of the peripheral and luminal proteins. The distribution profiles of the proteins across the density gradient were estimated from their tryptic peptides labeled by four-plex iTRAQ followed by strong cation exchange LC-MS/MS. To increase the number of fractions sampled, two four-plex labelings were performed per gradient. This allowed quantitation of seven fractions with one fraction common to the two labelings. The raw MS/MS spectral datasets from the two new LOPIT experiments and the two LOPIT experiments of Dunkley et al., (2006) were processed or reprocessed in parallel as described in materials and methods to produce protein abundance profiles. These analyses together identified and quantitated 2205 proteins. Confident subcellular localization requires observations in two or more independent biological samples. 1385 proteins achieved this threshold of significance (Supplementary Table S1). The number of proteins identified and quantitated was not very different between LOPIT experiments 1 and 2 vs 3 and 4, however, as expected, almost all of the novel proteins found only in the experiments using non-
carbonate-stripped membranes (LOPIT experiments 3 or 4) were without predicted TMDs (Supplementary Fig. S1).

Our next aim was to develop an approach to classify the identified proteins into organelle categories using the incomplete protein fractionation information from all eight four-plex labelings. It has been demonstrated by Trotter et al., (2010) that combining LOPIT datasets from different experiments improves the sub-cellular clustering resolution of proteins. However, a known issue with quantitation is that proteins are not observed in every MS experiment, due to the intensity-dependent manner in which mass spectrometry data is collected. Thus we expected to see only a minor proportion of proteins present and quantitated in all the experiments. Indeed, only 423 of the 2205 proteins identified and quantitated in these experiments were consistently detected in all eight iTRAQ labelings of the four experiments; the rest of the proteins had one or more missing measurements. These missing values can hinder subsequent multivariate analysis and classification algorithms, if not accounted for (Schafer, 1997). Imputation approaches have already been implemented in 2D-gel studies (Krogh et al., 2007; Pedreschi et al., 2008; Albrecht et al., 2010), and a few reports show the use of such algorithms in quantitative mass spectrometry (Du et al., 2008; Karpievitch et al., 2009). The imputation of missing values prior to applying the classification algorithm exploits the covariance of all available protein data. Common approaches to missing data imputation include estimation of the joint probability distribution of all variables drawn from the relationship among replicates (using Expectation-Maximization algorithm for instance) and sampling missing values from this model (Rubin, 1987). In multivariate datasets it can be difficult to specify a single joint distribution of all variables, so sequential sampling from a series of conditional predictive distributions is accepted as an approximation (Chen and Liu, 1996). A missing labeling in the LOPIT data consists of four values therefore it can be treated as a four-dimensional multivariate target variable, and in this case multivariate regression is a suitable choice for estimation. A PLS regression model was chosen as it allows the use of multivariate predictors and target variables even when the number of variables exceeds the number of cases (Wold et al., 1984). In this aspect the multiple imputations procedure adopted in this study is closely related to a Sequence of Regression Models technique (Raghunathan et al., 2001). The choice of PLS regression as a predictive model was validated by a test on a subset of the data with complete observations in all eight four-plex labelings (consisting of profiles for 423 proteins). The number of latent variables of the model was chosen by a seven-fold cross-validation scheme (Geladi and
Kowalski, 1986). The mean $R^2$ for a target variable was 0.87 with standard error of 0.0023. The predictive power was measured by a score ($Q^2$) where 0 is no better than chance and 1 is the theoretical maximum for a model that explains all the variation in the predicted data. The mean $Q^2$ of all classes was 0.802 with standard error of 0.0212, showing the validity of the approach.

We applied the data imputation approach to estimate the missing values on a data matrix consisting of partial distribution profiles along the density gradient of the 1385 proteins with two or more observations. The imputation scheme was validated by simulating random missing labelings and calculating the error of the imputation, and demonstrated an expected relative error of the magnitude of 10% to 16% of the real data value (Fig. 2). The generation of multiple imputed datasets was then achieved by a bootstrap approach (Efron, 1994). One hundred imputed datasets were generated by predicting the missing values with a series of PLS regression models where the training set was sampled with replication from the complete and already imputed cases. This procedure is considered to give better estimation of uncertainty than a single imputation of missing data with the best fit value or using a predictive model with a fixed set of parameters (Efron, 1994). For ease of presentation, the one hundred datasets were averaged into one matrix that yielded complete subcellular fractionation profiles on 1385 proteins in four experiments (Supplementary Table S2). This represents a twofold increase both in the amount of fractionation data for each protein and in the number of proteins over the experiments reported in Dunkley et al., (2006).

The averaged fractionation profiles were studied using principal component analysis (PCA) in order to visualize the clustering of protein profiles by their organelle localization (Fig. 3). In the reanalyzed experiments 1 and 2, the plots show that marker proteins of the same organelle clustered together (Fig. 3A) as seen in the earlier analysis of the MS/MS data (Dunkley et al., 2006). The two new data sets with the modified sample preparation also showed clear clustering of marker proteins (experiments 3 and 4, Fig. 3B). The combined data from the four experiments using the imputed fractionation values showed clear clustering similar to the separate datasets, again supporting the view that the data imputation was successful (Fig. 3C).

Proteins in the LOPIT datasets previously localized to the Golgi (35 proteins) and six proteins localized to the trans-Golgi network (TGN) were used as a part of the training set for the classification analysis and prediction of further Golgi/TGN associated proteins. The Golgi
apparatus and the TGN proteins were combined into one class because they have a similar fractionation profile. The rest of the training set was comprised of the proteins previously assigned to the PM (39 proteins), ER (53 proteins), vacuole (11 proteins), mitochondrion (92 proteins) and plastid (40 proteins), as well as a non-organelle (cytosolic/ribosomal) class (72 proteins) used to constrict the classifier. The training set is presented in Supplementary Table S2.

To classify proteins into compartments, a multivariate naïve Bayesian classification model with a kernel density estimate was applied to each of the one hundred imputed datasets. The choice of the naïve Bayesian model as a basis for multivariate classification was motivated by its particularly simple yet robust classification rule that uses a diagonal form of the covariance matrix. It has been reported to outperform a number of other popular classification models for multivariate biological data, including quadratic discriminant analysis and nearest neighbor classifier, and to yield better classification in situations where the sample values are correlated (Dudoit et al., 2002). Using the average posterior probability based on the naïve Bayesian classification models, 274 of the 275 membrane organelle training set proteins were assigned to the correct location. This indicates that the assignments using this approach are likely to be of high confidence.

The application of the classification procedure to the proteins with unknown localization resulted in a list of 197 proteins classified as Golgi/TGN-localized (Supplementary Table S2). As expected, most of the proteins reported to be Golgi-localized in Dunkley et al., (2006) lie within this set (85 out of 89). More than one-third (79 proteins) have not been previously identified to be Golgi-localized by a high confidence localization study (Supplementary Table S2, Supplementary references). This represents an increase of 120% over the Golgi proteome reported in Dunkley et al (2006). The same classification approach was applied to other organelle proteins, resulting in assignment of 304 proteins to the ER, 208 proteins to the PM, 213 proteins to the mitochondrion, 151 to the plastid and 37 proteins to the vacuole (Supplementary Table S2).

**The TMDs of proteins localized to different organelles demonstrate distinct properties**

TMDs of membrane proteins may harbor specific features that reflect the properties of the membrane in which they reside (Sharpe et al., 2010). Although in plants it has been shown that the TMD length has an effect on the localization of an artificial reporter protein (Brandizzi et al., 2002), there have been relatively few plant proteins localized, precluding
detailed investigation of differences in TMD properties of proteins in different organelles. The large number of proteins reliably localized by LOPIT here provides an opportunity to study whether plant proteins show organelle-specific TMD properties.

A subset of the proteins classified by the LOPIT approach was annotated as being putative single-span TMD proteins using the predictions from TMHMM (Table I; Krogh et al., 2001). The amino acid sequences of these proteins in and around their hydrophobic transmembrane span regions were analyzed similarly to Sharpe et al., (2010). This analysis refines the cytosolic hydrophobic start point of the TMD span according to a consistent heuristic function. The identification of hydrophobic start point allows aligning and computing different properties along the TMD sequence, such as average amino acid compositions, and corresponding derived parameters such as residue size and hydrophobicity. Also, measurements were made of the estimated extramembrane protein sequence length on the cytosolic side of the TMD from the same start point. To provide increased numbers of proteins for statistical analysis, the dataset was augmented by the inclusion of plant orthologs of the organelle-classified Arabidopsis proteins (Table I). The method for incorporating close ortholog sequences and removing familial sample bias was done as described previously for fungi and vertebrates (Sharpe et al., 2010). The sequence database used for the ortholog search was the Streptophyta phylum subset of the UniProt database. The mean hydrophobicity of amino acid residues of all the putative TMD sequences, at positions relative to the cytosolic ends of their hydrophobic cores, was calculated using the Goldman, Engelman and Steitz (GES) scale (Engelman et al., 1986) and plotted for the ER, Golgi and PM organelle classifications (Fig. 4A). The hydrophobic regions of the TMDs from the Golgi and ER were of similar maximum hydrophobicity, with this maximum extending over the same number of residues. In contrast for PM TMDs the region of maximum hydrophobicity extended for a further five residues (Fig. 4A), similar to the situation observed in fungi and vertebrates (Sharpe et al., 2010). To demonstrate that positional averages did not obscure any differences, we plotted the frequency of hydrophobic length, defined with an intermediate GES scale threshold of 1 kcal/mol (Fig. 4B). The distribution plot shows distinct profiles for the three compartments, with the mean TMD hydrophobic span lengths of 21.5 residues for the ER, 22.7 for the Golgi and 25.4 for the PM (Fig. 4B). We tested the dependence of the hydrophobic length output on the GES scale threshold used to define TMD ends, and concluded that the difference between PM, ER and Golgi is relatively unaffected by the threshold chosen, suggesting that the difference is robust (data not shown).
Since residue volumes are distributed asymmetrically in PM and Golgi TMDs of fungi and vertebrates (Sharpe et al., 2010), the mean residue volume was calculated along the plant TMD sequences (Fig. 4C). The TMD regions of different organelles showed no significant differences regarding the amino acid volume close to the cytosolic edge (positions 1 to 10). However, in the luminal/exoplasmic leaflet region, starting from position 12 there are clear differences: the PM TMDs have a significant drop in the mean residue volume, whereas Golgi and ER remain similarly high throughout the TMD (Fig. 4C). To investigate this in more detail, we analyzed the frequency of each amino acid at each position through the TMD. The amino acid compositions show differences between the TMDs from different organelles (Supplementary Fig S2). Some of these differences are clearly associated with the differences in mean residue volume. In particular, the larger hydrophobic residues (Leu, Phe and Tyr) are relatively depleted in the outer leaflet portion of the PM TMDs compared to Golgi and ER, whilst the smaller hydrophobic residues (Ala, Ile and Val) show the opposite trend (Figs. 4E and 4F). In addition to these differences in residue size, there appear to be other organelle-specific trends. For example, the PM TMDs have enriched abundance of the basic residues on the cytosolic side, and increased occurrence of Cys in the interfacial region of the inner leaflet (perhaps reflecting increased palmitoylation). The Golgi TMDs are characterized by lower occurrence of Lys in the outer leaflet region, whereas the ER TMDs have increased occurrence of Met in the interfacial region of the outer leaflet (Supplementary Fig. S2).

The length of the cytosolic tail of the single-span TMD proteins also displayed specific variations for the three organelles. Over 80% of the Golgi apparatus proteins have a cytosolic tail of fewer than 45 amino acid residues. Almost as frequently, 70% of ER proteins have similarly short cytosolic tails; whereas this observation holds for fewer than 22% of PM proteins (Fig. 4D). This is to be expected, since the transmembrane PM proteins (unlike the ER and Golgi proteins) are likely to perform more functions on the cytosolic side.

**Novel putative GTs are identified in the Golgi proteome**

Many of the proteins (22%) identified as Golgi-localized were annotated in the TAIR10 database as proteins of unknown function. Since a large number of the proteins in the Golgi apparatus are likely to be involved in glycosylation (Dupree and Sherrier, 1998), we investigated whether any of these proteins show predicted structural homology to proteins of known function and structure. Currently, 3-D structural information is available for 39 of the

---

Copyright © 2012 American Society of Plant Biologists. All rights reserved.
91 classified CAZy GT families. We analyzed the sequences of all the 197 Golgi proteins using HHpred, an alignment tool based on the pair-wise comparison of Hidden Markov Model (HMM) profiles (Soding et al., 2005). HHpred builds a HMM profile from a query amino acid sequence and compares it with a database of HMMs representing annotated protein families (e.g. Pfam, SMART, CDD) or domains with known structure (PDB, SCOP) and provides greater sensitivity and selectivity in finding remote homologs than other sequence-profile based methods such as PSI-BLAST (Soding, 2005). The analysis of all the Golgi-localized proteins identified 65 of 197 as having a probable fold related to known GTs; of these 41 are classified in CAZy and 24 proteins were identified as putative GTs not classified in CAZy families (see Table II). In addition, the DUF579 family as well as three other unrelated proteins showed significant similarity to a conserved S-Adenosyl Methionine (SAM) binding methyltransferase (MT) fold (Supplementary Fig. S3). To obtain an overview of the Golgi proteome, all the 197 proteins were assigned to functional groups and these are shown in Figure 5 and Supplementary table S2.

DISCUSSION

Replication of quantitative proteomic experiments and data integration allow generation of larger quantitative datasets

The major difficulty in analyzing the Golgi proteome for many years was the purification of the organelle, because of its low abundance in the cell and its similar density to other cellular membranes (Mo et al., 2003). Furthermore, the Golgi apparatus contains proteins en route to other organelles because of anterograde and retrograde trafficking in the endomembrane system. Thus, even a pure Golgi preparation will not solely contain Golgi resident proteins. The LOPIT method was developed to overcome these obstacles and is based on De Duve’s observations that a given organelle is not solely present in one particular density gradient fraction but rather forms a distribution profile along the gradient (De Duve, 1971; Dunkley et al., 2004). LOPIT describes the steady state locations of proteins, thus allowing more permanent residents of the Golgi to be distinguished from cargo. Here, we reanalyzed previous datasets with an updated proteome database and combined it with two new LOPIT experiments. To overcome problems of multivariate statistical analysis of the incomplete observations across all these proteomic datasets, we introduced a statistical imputation
approach. This allowed us to identify the subcellular localization of 1110 proteins with a high confidence. This is more than twice as many proteins as localised in Dunkley et al., (2006).

Characteristics of plant ER, Golgi and PM protein membrane domains

Our collection of high quality datasets of protein localization allowed us to investigate properties of the single-span TMDs in different organelles. TMDs of membrane proteins are likely to have features that reflect the properties of the membrane in which they reside (Sharpe et al., 2010), and these features may even influence their trafficking. It has been shown that TMDs in fungi and animals have organelle-specific properties, including a difference in TMD length between the early and late parts of the secretory pathway (Sharpe et al., 2010). A similar analysis of plant organelle proteins, based on our newly expanded LOPIT data, indicates that there are comparable differences in plant TMD length and amino acid composition (Fig. 4). The length of the TMD region showing the maximum average hydrophobicity is longer for the PM than for the Golgi and ER proteins by five residues. This difference is comparable to that observed for vertebrates (six residues), with fungi showing an even larger difference (nine residues) (Sharpe et al., 2010). The average amino acid volume also shows a distinction between the organelles, with smaller residues being preferred in the exoplasmic leaflet of PM TMDs. Again, this is qualitatively similar to the situation in fungi and vertebrates, but this time the difference is more marked than in vertebrates, although again smaller than in fungi. The resulting plots and data of amino acid compositions, properties and abundance were submitted to http://www.tmdsonline.org and the differences in TMD properties described here could be used to improve the targeting prediction algorithm for single-span transmembrane proteins.

It has been suggested for fungi and vertebrates that these TMD differences reflect organelle-specific physical properties of the bilayer, with the PM being thicker with a particularly well ordered outer leaflet (Sharpe et al., 2010). As in fungi and vertebrates, the PM of plants is enriched in sterols and sphingolipids (Hartmann, 1998; Pata et al., 2010). The precise structure of these lipids varies between the different clades, but the plant sterols and the glycosyl inositol-phosphoceramides are believed to be abundant in plant PMs, perhaps even accounting for the large majority of the lipid molecules in the outer leaflet (Sperling et al., 2005). The ordering effect of the rigid sterols and the saturated, or trans double-bonded, acyl chains of the sphingolipids would help exclude toxins and other undesirable molecules as well as having the effect of thickening the bilayer. Proteins would need to adapt to this
different bilayer, and it may also be that the TMD differences could contribute to protein sorting in the secretory pathway in plants, as in the other two clades (Brandizzi et al., 2002; Wang et al., 2008).

**Identification of the cell wall synthesis machinery in the Golgi apparatus**

Our rationale for extending the cataloging of the most abundant proteins in the Golgi apparatus is that within the identified Golgi proteins, most of the proteins involved in primary cell wall polysaccharide synthesis would be present, particularly proteins in pathways of synthesis of the more abundant polysaccharides. The Arabidopsis liquid-grown callus cell wall is a typical primary cell wall, and is composed of abundant polygalacturonan, rhamnogalacturonan I and xyloglucan with minor amounts of type II arabinogalactan, extensin, glucuronoroarabinoxylan, rhamnogalacturonan II and glucomannan (Goubet et al., 2002; Handford et al., 2003; Manfield et al., 2004; Barton et al., 2006; PD unpublished observations). We therefore expected to identify proteins involved in the synthesis of these polysaccharides.

One-fifth (41) of the 197 identified Golgi proteins are putative or confirmed GTs described in CAZy, strongly supporting the view that the callus Golgi apparatus is rich in GTs and is active in polysaccharide synthesis and protein and lipid glycosylation (Fig. 5). For many of these CAZy classified GTs, we provide here the first evidence for Golgi localization, including the localization of two of the newly CAZy classified GT90 enzymes. Xyloglucan is the main hemicellulose of dicot primary cell walls (Scheller and Ulvskov, 2010). We localized most of the GTs shown to be involved in xyloglucan synthesis to the Golgi (XT1, XT2, XXT5, MUR3, AtCslC4, Madson et al., 2003; Cocuron et al., 2007; Cavalier et al., 2008; Zabotina et al., 2008). Consistent with high pectin synthesis in the callus cells, we identified GAUT1, GAUT7 and many other GAUT family members in GT8 (Sterling et al., 2006; Atmodjo et al., 2011), as well as ARAD1 and XGD1 in GT47 involved in arabinan and xylogalacturonan synthesis respectively (Harholt et al., 2006; Jensen et al., 2008). We also detected enzymes likely involved in primary cell wall glucuronoroarabinoxylan synthesis: IRX10-L, IRX14, and two IRX15/DUF579 homologs (Brown et al., 2009; Wu et al., 2010; Brown et al., 2011; Jensen et al., 2011), and GUX3, a homolog of the GUX1 and GUX2 xylan glucuronosyltransferases (Mortimer et al., 2010). Proteins of extensin glycosylation (RRA3 and XEG113, Velasquez et al., 2011) and possibly glucomannan synthesis (CslD2 and CslD3, Yin et al., 2011) were identified as well, indicating that LOPIT also detects...
proteins synthesizing relatively minor cell wall components. Two subunits of the cellulose synthase were localized to the Golgi/TGN at steady state, in contrast to four callose synthases that were found in the plasma membrane. The remaining CAZy classified Golgi GTs are involved in N-linked glycosylation or are of unknown function. A very recent paper from Heazlewood and colleagues (Parsons et al., 2012) describes a study of the Arabidopsis Golgi proteome using different membrane separation techniques and MS methods. Many of the proteins localized in the Golgi in this work were also found in their Golgi preparations.

Identification of putative Golgi GTs

To address the question of whether further identified Golgi proteins may be involved in glycosylation, we used comparative structural modeling to search for protein homologs with known function. Since characterized sugar-nucleotide binding Golgi GTs have one of two folds (GT-A or GT-B), unknown GTs are likely to have a similar 3-D structure (Unligil and Rini, 2000; Lairson et al., 2008). We applied HHpred, a sensitive structural prediction and remote homology detection algorithm, to the identified 197 Golgi proteins. In this way, we identified an additional 24 putative GTs in 12 sequence-related families (Table II). This bioinformatic prediction of GT-like folds is supported by our demonstration that they are Golgi residents. There are in total 96 proteins in these 12 putative Golgi GT families encoded in the Arabidopsis genome (Supplementary Table S3). CAZy GT families have at least one member biochemically characterized (Coutinho et al., 2003). We detected 41 putative or confirmed CAZy GTs by LOPIT (from 13 CAZy families, having a total of 230 members). Comparing the numbers of putative GTs in these two categories found in this study, we speculate that perhaps a third of Arabidopsis Golgi GTs are not yet assigned to CAZy GT families because of the lack of biochemical characterization (Fig. 5). Egelund et al., (2004) and Hansen et al., (2009) also used structural bioinformatic approaches to identify some of these putative GTs. One of their identified families has since been shown to have GT activity and was added to the CAZy database as GT77 (Egelund et al., 2007; Gille et al., 2009).

KOBITO1 has earlier been reported to be necessary for cellulose synthesis in primary cell walls and localized to the PM of elongating root cells and in an internal compartment (Pagant et al., 2002). Here we demonstrate the Golgi/TGN localization of this protein in callus (Supplementary Table S2). This is consistent with the similar finding that the cellulose synthases are also localized in the Golgi/TGN at steady state (Dunkley et al., 2006; Supplementary Table S2), but during cell elongation are targeted to the PM where they are
functionally active (Paredez et al., 2006; Crowell et al., 2009; Wightman et al., 2009). The HHpred alignment showed homology of KOBITO1, and its two homologs (At2g41451 and At3g57200), to the 3-D structures of GT2, and they are likely to have a GT-A fold. We name this group of three proteins GT92R because the putative GT domain is most closely related to GT92 (Table II, Supplementary Fig. S4). A recent study also identified KOBITO1 as a putative GT and proposed a function in a carbohydrate metabolic process essential for proper plasmodesmata closure (Kong et al., 2012).

The GT65R is a large family, and consists of 35 members containing DUF246 domains in Arabidopsis, eight of which were shown in this study to be Golgi-localized. This family has already been predicted to be a putative GT family (Egelund et al., 2004; Hansen et al., 2009), and in the Pfam database this group has been recently fused into PF10250 which contains metazoan protein O-fucosyltransferases. Since the GT65R family shows structural homology to GT65, it is likely to have a GT-B fold. However, the sequence similarity to GT65 is low, suggesting the activity may be different, and only one of its members (At1g52630) is currently in the CAZy database, in the group of "non-classified" GTs (GTNC). Two genes have been partially characterized: RHS17 (root hair specific 17) and EDA30/TGD1 (embryo sac development arrest 30 / tube growth defect 1). The RHS17 gene has a root hair-specific expression pattern and may play a role in root hair growth and morphogenesis (Won et al., 2009). The eda30 mutant fails in fusion of polar nuclei in the embryo sac (Pagnussat et al., 2005), another mutant in the same gene, tgd1, had defective pollen tube growth (Boavida et al., 2009).

The GT68R-A and GT68R-B families in Arabidopsis consist of three and two members, respectively. There are some distant similarities to GT65R (DUF246), since they show structural homology to the GT-B fold GT65 family. SUB1, a member of the GT68R-A, has previously been suggested to be involved in the cryptochrome signaling pathway as a calcium-binding protein in the ER (Guo et al., 2001). Our results of SUB1 localization to the Golgi apparatus suggest that the protein fusion of SUB1 to β-glucuronidase was possibly mislocalized.

The GT14R consists of 23 proteins with DUF266 domains in Arabidopsis, one of which was localized in our proteomic experiments. This family has already been predicted to be a putative GT family (Egelund et al., 2004; Hansen et al., 2009) and most of the members added to the CAZy GTNC. It was shown that this domain is distantly related to GT14, likely
possessing the GT-A fold, but lacks the DxD motif (Hansen et al., 2009). GT14 contains the characterized N-glycan beta-1,6-N-acetylglucosaminyltransferase of animals, and there are eleven GT14 in Arabidopsis, but there are no characterized members in plants. In rice, the bc10 mutant with a loss of function of a DUF266 protein has an altered cell wall phenotype, suggesting that BC10 acts as a Golgi GT involved in cell wall synthesis (Zhou et al., 2009).

Two novel families have putative GT8 like structures: GT8R-A (DUF616) and GT8R-B. In Arabidopsis there are eight GT8R-A proteins, containing the DUF616 domain, and one was found to be Golgi-localized in this study. The DUF616 is a domain present in bacterial and plant proteins. In several bacterial proteins this domain is fused with GT2 domains (e.g. Uniprot accessions Q2GAF2 and B3EF89). DUF616 shows highest structural homology to the GT-A fold GT43 family and to the GATL clade of GT8 in Arabidopsis (Table II). Several Arabidopsis GT8R-A genes are co-expressed with genes encoding pectin biosynthetic enzymes (http://atted.jp, Obayashi et al., 2007); hence a possible function of this class of proteins could be in the synthesis of pectic polysaccharides. We have identified and localized two of the four GT8R-B family members in Arabidopsis. Similarly to GT8R-A, the GT8R-B family shows homology to GT43 and to the GATL clade of GT8 in Arabidopsis (Table II).

The GT27R family in Arabidopsis comprises of 11 proteins, having a DUF707 domain and is restricted to plants and prokaryotes. Two GT27R proteins have been localized in our study to the Arabidopsis Golgi apparatus. The domain shows highest structural homology to the GT-A fold GT27 and to the GT2 family in Arabidopsis (Table II).

The GT13R protein (AT5G12260) is a singleton. Its protein sequence suggested structural homology with two GT families: the N-terminal domain to GT13, and the C-terminal domain to GT77. Therefore, this protein appears to be a bi-modal protein with two GT domains, similar to some heparan sulfate synthase proteins in animals. These fusions of GT47 and GT64 domains transfer alternately β-1,4-glucuronyl and α-1,4-N-acetylglucosaminyl residues (Sugahara and Kitagawa, 2002). It is possible that the Arabidopsis GT13R performs a similar alternating transfer of two different sugar residues (GT13 has an inverting and GT77 a retaining mechanism), and hence it could be involved in biosynthesis of a polymer with alternating sugar backbone, such as rhamnogalacturonan I.

The GT4R and GT75R proteins constitute two further putative GT families in Arabidopsis. We have identified and localized both GT4R and one of the two GT75R proteins encoded in the genome. The homology modeling shows that the DUF288 domains in GT75R are related
to the reversibly glycosylated proteins/ UDP-arabinose mutase (GT75), and are structurally close to GT2, with a GT-A fold (Table II).

Several glycoside hydrolases (GHS) were also detected in this Golgi proteome study: mannosidases, belonging to GH38 and GH47, presumably responsible for the maturation of N-glycans. However, we did not detect any Golgi localized GHS that may be involved in remodeling of cell wall polysaccharides.

**Transporters and other multiple membrane spanning proteins**

The Golgi apparatus transports various metabolites for luminal polysaccharide synthesis, including sugar nucleotides, SAM and probably Acetyl-CoA (Reyes and Orellana, 2008). Three families have been identified that fit the criteria of type III membrane proteins and are likely transporters: NST/TPT (nucleotide sugar transporter / triose phosphate translocator), DUF791 and EMP70.

Six members of the NST/TPT family were identified (of 53 members proposed by Reyes and Orellana, 2008). The members of this family possess six to ten transmembrane helices (Knappe et al., 2003) and most of them have a DUF250 domain. Four of the identified proteins in this study belong to the KD clade and one to the KT clade of PTh (phosphate translocator homologous), as annotated in Knappe et al., (2003). Another nucleotide sugar transporter is the uncharacterized (At4g35335), which has the highest similarity to UTr6. One other uncharacterized transporter-like protein was identified, At4g39840, which contains five-to-six TMDs.

There are four DUF791 family members in Arabidopsis and two of those were identified to be Golgi-localized in this study. This family is part of the major facilitator superfamily and they show homology to sugar proton symporters. Members of DUF791 are predicted to have 12 TMDs.

The endomembrane protein 70 (EMP70) family is predicted to have a large luminal domain after the signal sequence and nine conserved putative TMDs, giving the name Nonaspanin domain. Twelve proteins are encoded by the Arabidopsis genome, 10 of which were detected and localized to the Golgi by our approach. It has been shown that a yeast ortholog of EMP70 is important for early endosome-to-vacuole trafficking and vacuolar protein sorting (Aguilar et al., 2010). However, Arabidopsis EMP70 members have been shown to be localized exclusively to the Golgi (Dunkley et al., 2006, Gao et al., 2012). This family is quite
conserved through all eukaryotes and this suggests that members of EMP70 play important roles in secretion and/or Golgi functioning. The EMP70 proteins may function as channels or transporters.

**Sugar nucleotide metabolism**

Several enzymes catalyzing the nucleotide sugar interconversion reactions were identified with this approach as Golgi residents: GAE1 and GAE5, UXS1, UXS2, UXS4. The presence of UDP-Gal and UDP-xylose synthesis enzymes is consistent with the high biosynthetic activity of xyloglucan and pectin. We also identified a previously undescribed putative epimerase/isomerase (At3g56820). The structure-based homology search against the CDD database showed it is an rmlC-like protein. Proteins in this family have one or both of C3 and C5 epimerase activities, such as bacterial dTDP-4-keto-6-deoxyglucose 3,5-epimerase in dTDP-rhamnose synthesis. Another group of proteins that includes activities of nucleotide sugar interconversions is the Kelch motif containing family; two such proteins were identified as Golgi residents (At1g51540 and At3g27220).

**Putative S-AdenosylMethionine (SAM) dependent MethylTransferases (MTs)**

There were a large number (24) of putative MTs detected in the Golgi by LOPIT. The functions of this type of Golgi enzymes are largely unknown. It is likely that most of these enzymes are involved in methylation of the glycans. MTs are an extensive class of enzymes that share little sequence identity, but contain a well conserved structural fold. Many crystal structures of SAM-dependent MTs have been determined and they define the core SAM-MT fold as consisting of a central seven-stranded β-sheet that is flanked by three α-helices per side of the sheet. The SAM binding region is localized to the N-terminal part and comprises of the first three N-terminal β-strands separated by α helices (Martin and McMillan, 2002). This conservation allowed us to discover distantly related proteins in the MT superfamily.

We identified 19 Golgi proteins containing the DUF248 domain. They have collectively been annotated previously as a putative MT family, with one member, QUA2, being involved in homogalacturonan methylation (Mouille et al., 2007; Ralet et al., 2008). The DUF579 domain shows high predicted structural similarity to several MTs (Supplementary Fig. S3). Two DUF579 family members have been shown to be involved in xylan biosynthesis – IRX15 and IRX15L (Brown et al., 2011; Jensen et al., 2011), but it is still unclear how they are involved. Three additional putative MTs assigned to the Golgi are encoded by the
singleton At3g16200 and the two closely related At3g49720 and At5g65810. These are only
distantly related to both DUF248 and DUF579.

CONCLUSIONS

The localization of over 1100 plant membrane proteins will be an invaluable resource for the
community and allow in depth analysis of many organelle properties and functions. For
example, the identification here of differences in TMD properties between the ER, Golgi and
PM proteins will allow better prediction of novel proteins localization as well as facilitate
studies of protein targeting. The currently identified GTs are insufficient to synthesize all the
cell wall carbohydrate structures produced by plants. We identified 12 biochemically
uncharacterized Golgi putative GT families, extending substantially the number of targets for
functional studies which are now underway in our laboratory. We believe that these results
will greatly contribute to the understanding of plant cell wall polysaccharide biosynthesis.

MATERIALS AND METHODS

Growth of plant material, organelle separation, protein digestion, labeling and MS
analysis

We used the established Arabidopsis thaliana liquid-grown callus for the proteomic analysis.
The conditions of growth were as described in Prime et al., (2000). Organelle separation was
achieved from homogenized Arabidopsis cell culture using density centrifugation as
described before (Sadowski et al., 2006). For the peripheral LOPIT experiment the membrane
fractions obtained from the density centrifugation were not treated with sodium carbonate.
Proteolysis was achieved with trypsin and the peptides were labeled with four-plex iTRAQ,
separated by strong cation exchange chromatography and analyzed on a AB Sciex QSTAR
Q-ToF LC-MS system as described previously (Dunkley et al., 2004; Dunkley et al., 2006).

Data processing – integration of four data sets and missing values imputation

The raw MS/MS data files obtained from the QSTAR were processed using the wiff2DTA
software to generate centroided and uncentroided peak lists (Boehm et al., 2004) containing
m/z and intensity information for each product ion spectrum. The centroided peak lists were
analyzed by MASCOT 2.2 (Matrix Science, London) and queried against the TAIR8 non-
redundant protein database (27,234 sequences) using the following modifications: fixed: MMTS (Cys), variable: iTRAQ four-plex (Lys), iTRAQ four-plex (N-terminal), iTRAQ four-plex (Tyr), oxidation (Met). The ions score significance threshold was set to 0.05, MS tolerance was set to 1 Da and the MS/MS tolerance to 0.8 Da. Each centroided peak list was also searched separately against a decoy reversed version of the database in order to determine the MASCOT peptide score resulting in a false identification rate of <1% in each LOPIT experiment. Hence the minimum peptide MASCOT scores for proteins identified with two or more peptides were: 32 for LOPIT experiment 1, 35 for LOPIT experiment 2, 33 for LOPIT experiment 3 and 33 for LOPIT experiment 4. The thresholds for identification of proteins with single peptides were 47, 51, 52, and 50 for experiments 1 to 4 respectively. Proteins identified by two or more distinct peptides in at least one experiment were retained. The XML files created from the integrated dataset were exported to the PRIDE database (Vizcaino et al., 2010). The uncentroided peak lists were processed with i-Tracker software in order to calculate the normalized iTRAQ reporter ion areas (Shadforth et al., 2005). The Genome Annotating Proteome Pipeline system (Shadforth et al., 2006) was used to link identification information with quantitation data in a MySQL database. Peptides were assigned to proteins only if at least three of their reporter ion peaks were above a threshold of 15 counts and if they had a MASCOT score of at least 20 (Dunkley et al., 2006). In addition, peptides matching multiple proteins were not used for protein quantitation. The quantitative data obtained from the four reporter ions of the iTRAQ label was corrected for the isotope impurities and normalized such that the four reporter ions sum to one (Fig. 1). Quantitation data from multiple peptides from the same protein were averaged to provide a fractionation profile for identified proteins. Proteins identified and quantitated in two or more experiments were retained. The normalized four-plex datasets were concatenated in one matrix presented in Supplementary Table S1.

**Multivariate data analysis**

Protein profiles that were present in only one experiment were discarded (See Data pre-processing on Fig. 1). The protein profiles that had missing four-plex data (from one up to six missing four-plex data out of eight) were imputed by a PLS regression model (Geladi and Kowalski, 1986). The imputation was performed sequentially; first the entire dataset was divided into subgroups according to the number and position of missing four-plex data. For each subgroup, the missing labeling was estimated by a PLS regression model built using proteins where all labelings were present. The estimated labeling was then considered
‘known’ and used as a predictor for building subsequent regression models. The process was performed first for the protein profiles having one four-plex labelings missing, and then repeated for those with two missing labelings. This process was iterated until all the missing values were imputed. The validation of this method was performed using a cross-validation approach. One non-missing four-plex was randomly removed from each protein with up to five missing labelings and the PLS regression method was applied to this new dataset in order to predict known and unknown missing values. The predicted four-plexes were then compared to the known values and the prediction relative error was calculated. This process was repeated 1000 times, and expected error (summarized by subgroups of proteins with the same number of missing values) was estimated by a mean error and its 95% confidence interval.

The list of protein IDs with known subcellular localization was manually curated from published research papers, with the assistance of the TAIR (www.arabidopsis.org; Lamesch et al., 2012) and SUBA databases (http://suba.plantenergy.uwa.edu.au; Heazlewood et al., 2007). The training set was composed of proteins localized by methods that were LOPIT orthogonal and considered high confidence (fluorescence microscopy imaging, immunolocalization, or, in cases of proteins localized to the plastid or mitochondrion, sequence homology and protein import assays). This way we gathered a list of 348 protein IDs distributed in seven classes (41 proteins belonging to the Golgi/TGN class, 53 to the ER, 39 to the PM, 11 to the vacuole, 92 to the mitochondrion 40 to the plastid, plus 72 to a decoy class consisting of ribosomal and cytosolic proteins; Supplementary Table S2). Their corresponding quantity profiles were used as a training set for the classification model. The multivariate classification rule with diagonal covariance matrices (known as naïve Bayesian classifier) was used to build a multivariate classification model. The prior probability for a particular protein to be assigned to one of six classes was set uniformly. The class posterior distributions were modeled in a non-parametric way using a Gaussian smoothing kernel.

In order to account for the uncertainty of classification, originating from the missing value imputation step, a multiple imputation approach was undertaken as described previously (Rubin, 1987). Exactly 100 variants of an imputed dataset were generated using the bootstrapped sequential imputation procedure described above in order to explore the variation in the prediction. Classification models were built for each of the 100 generated datasets. The average posterior probability was calculated to reflect the variability of the class probability distributions and the imputed data. The performance of the classifier was
evaluated using a leave-one-out cross-validation scheme where the localization of one of the proteins from the training set was predicted using the remaining training set proteins. The proteins in the training set were correctly assigned in 98% of these models. Proteins were finally assigned using organelle-specific thresholds based on the average posterior probability of assignments of the training set proteins, and the predicted classes are presented in Supplementary Table S2. The data processing was done using the Matlab Statistical Toolbox (Mathworks Inc).

**TMD analysis**

The single-span TMD proteins from the LOPIT classified set were identified using a combination of sequence filtering, employing SignalP version 4.0 (Petersen et al., 2011), TMHMM (Krogh et al., 2001) and GPI anchor prediction (Borner et al., 2002). In the first pass only protein sequences were accepted, according to the TMHMM prediction of the TM spans, if the protein appears to have only one TM span or if it appears to have two, but the first span is at the N-terminus (and so could be a signal peptide). If a TM span was at the N-terminus and predicted to be a signal sequence, according to SignalP, then the protein was excluded if this was the only TM span, but accepted if this was the first of two TMHMM predictions. Any sequence that passed these tests was subject to the "GPT" method of GPI anchor prediction used in Borner et al., (2002), which excluded some putative GPI sequences on the basis of a C-terminal hydrophobic signal peptide.

In order to augment the organelle classified Arabidopsis sequences, close orthologs in related organisms were identified in the Streptophyta clade within the UniProt database. The expanded lists of protein sequences were reduced for redundancy and analyzed for their cytosolic extra membrane length and TMD sequence properties: amino acid composition, hydrophobicity, and residue volume, as described in Sharpe et al., (2010).

**Protein structure prediction**

The remote homology detection server HHpred (www.toolkit.tuebingen.mpg.de/hhpred) was used to search for homologs to the uncharacterized protein families (Soding et al., 2005). All searches were conducted with the default settings: three HHblits iterations were performed in local alignment mode to build up a multiple sequence alignment from the input sequence. The algorithm was used to search the Protein Data Bank for 3-D structures, Pfam and CDD
for primary structures and the Arabidopsis protein database for determining the closest CAZy GT family to the query.

ACKNOWLEDGEMENTS
We thank Zhinong Zhang for maintaining the Arabidopsis callus lines, Svenja Hester and Julie Howard for assistance with mass spectrometry, Hayley Sharpe for providing part of the python script, David Burke for useful discussions on protein structural prediction, Bernard Henrissat, Pavel Shliaha and Laurent Gatto for critical reading and comments on the manuscript.
REFERENCES


Caffall KH, Mohnen D (2009) The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. Carbohydr Res 344: 1879-1900


Mapping the Arabidopsis organelle proteome. Proc Natl Acad Sci U S A 103: 6518-6523


glycosylated cell wall proteins are essential in root hair growth. Science 332: 1401-1403


membrane protein, is required for cell-wall biosynthesis in rice (Oryza sativa L.).
Plant J 57: 446-462
### Table I. Summary of the number of organelle proteins used in the LOPIT approach and in the TMD analysis

<table>
<thead>
<tr>
<th>Organelle</th>
<th>LOPIT Known localization (training set)</th>
<th>Assigned localization</th>
<th>TMD analysis Single-span TMD proteins used in the analysis</th>
<th>With orthologs (Streptophyta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Golgi/TGN</td>
<td>41</td>
<td>197</td>
<td>83</td>
<td>321</td>
</tr>
<tr>
<td>ER</td>
<td>53</td>
<td>304</td>
<td>78</td>
<td>395</td>
</tr>
<tr>
<td>PM</td>
<td>39</td>
<td>208</td>
<td>23</td>
<td>136</td>
</tr>
</tbody>
</table>
Table II. The 12 putative GT families of the 24 proteins identified as Golgi-localized and having predicted structural homology to CAZy classified GTs. For a full list of Arabidopsis GT-like proteins see Supplementary table S3.

<table>
<thead>
<tr>
<th>Gene family name</th>
<th>No. of proteins found in this study</th>
<th>Structural prediction and remote homology detection</th>
<th>Probability (%)</th>
<th>Best Arabidopsis CAZy classified GT hit</th>
<th>Probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT92R</td>
<td>1</td>
<td>1g8h [GT2, G-T-A fold, inverting]</td>
<td>97.8</td>
<td>AT5G44670 [GT92]</td>
<td>99.3</td>
</tr>
<tr>
<td>GT65R</td>
<td>9</td>
<td>3zy2 [GT65, G-T-B fold, inverting]</td>
<td>100</td>
<td>AT3G03270 [GT65]</td>
<td>99.9</td>
</tr>
<tr>
<td>GT68 R-A</td>
<td>1</td>
<td>3zy2 [GT65, G-T-B fold, inverting]</td>
<td>99.6</td>
<td>At1g52630 [GT68]</td>
<td>98.8</td>
</tr>
<tr>
<td>GT68 R-B</td>
<td>2</td>
<td>3zy2 [GT65, G-T-B fold, inverting]</td>
<td>99.8</td>
<td>At1g53770 [GT68]</td>
<td>99.8</td>
</tr>
<tr>
<td>GT14R</td>
<td>1</td>
<td>2gak [GT14, G-T-A fold, inverting]</td>
<td>100</td>
<td>At2g37585 [GT14]</td>
<td>100</td>
</tr>
<tr>
<td>GT75R</td>
<td>1</td>
<td>2d0j [GT43, G-T-A fold, inverting]</td>
<td>69.9</td>
<td>At1g50750 [GT75]</td>
<td>100</td>
</tr>
<tr>
<td>GT8 R-A</td>
<td>1</td>
<td>1g9r [GT8, G-T-A fold, retaining]</td>
<td>84.2</td>
<td>At3g50760 [GT8]</td>
<td>83.6</td>
</tr>
<tr>
<td>GT8 R-B</td>
<td>2</td>
<td>1g9r [GT8, G-T-A fold, retaining]</td>
<td>59.7</td>
<td>At3g50620 [GT8]</td>
<td>71.3</td>
</tr>
<tr>
<td>GT27R</td>
<td>2</td>
<td>2ffu [GT27, G-T-A fold, retaining]</td>
<td>95</td>
<td>At1g20575 [GT2]</td>
<td>96.4</td>
</tr>
<tr>
<td>GT4R</td>
<td>2</td>
<td>3c48 [GT4, G-T-B fold, retaining]</td>
<td>100</td>
<td>At3g45100 [GT4]</td>
<td>100</td>
</tr>
<tr>
<td>GT13R</td>
<td>1</td>
<td>1f08 [GT13, G-T-A fold, inverting]</td>
<td>100</td>
<td>At3g38240 [GT13]</td>
<td>100</td>
</tr>
<tr>
<td>GT41R</td>
<td>1</td>
<td>3tax [GT41, G-T-B fold, inverting]</td>
<td>100</td>
<td>At3g11540 [GT41]</td>
<td>100</td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

Figure 1. The LOPIT experimental design and the data pre-processing workflow

Organelle separation is obtained by density gradient ultracentrifugation. Four LOPIT experiments were performed, each consisting of two four-plex iTRAQ labelings from seven density gradient fractions. One fraction is common in both four-plex labelings. The peptide abundance values from the MS/MS quantitation of the iTRAQ reporter ions are collapsed to a protein level for each four-plex labeling and are presented as color-coded values. The missing value imputation was performed on protein profiles that were present in at least two experiments.

Figure 2. Evaluation of the missing values imputation approach

The mean expected error of imputed data and its 95% confidence intervals (CI) are presented. The plot shows the result of 1000 simulation runs of the imputation procedure.

Figure 3. Multivariate analysis of the LOPIT data

(A) PCA plot of 728 protein profiles from the LOPIT experiments 1 and 2; (B) PCA plot of 717 protein profiles from the LOPIT experiments 3 and 4; (C) PCA plot of the 1385 protein profiles of the combined datasets (LOPIT experiments 1, 2, 3 and 4) with imputed missing values. The profiles of marker proteins with previously reported subcellular localisation (training set proteins) are colored in the PCA plots based on the organelle of residence.

Figure 4. Analysis of the TMD characteristics of single-span membrane proteins classified as localised to the ER, Golgi/TGN and PM

(A) The mean hydrophobicity (GES scale) of the residues at each position along the aligned TMDs relative to the cytosolic edge. The GES hydrophobicity values represent the free energy of partitioning from water into a hydrophobic environment; (B) The distribution of TMD lengths. The exoplasmic ends of the TMD were defined using the hydrophobicity scanning algorithm as for the cytosolic ends; (C) The mean values for the residue volume at each position along the TMDs; (D) Cumulative distribution of the cytosolic extramembrane lengths; (E) Analysis of the cumulative abundance of leucine, phenylalanine and tyrosine along the TMDs; (F) Analysis of the cumulative abundance of alanine, isoleucine and valine along the TMDs.
Figure 5. Functional class assignment of the 197 Golgi-localized proteins

The classification is based on the TAIR annotation for the characterized proteins and the HHpred output predictions for the uncharacterized proteins, as indicated in Supplementary table S2. The numbers of proteins per class are indicated in brackets.
Hydrophobicity

A

TM Hydrophobic Length Distribution

B

Amino acid volume

C

Cytosolic extramembrane length

D

Composition of Leu+Phe+Tyr

E

Composition of Ala+Ile+Val

F