Supplemental material

**Missing values imputation**

In the table below we report the percentage rates of MVs per group of samples:

<table>
<thead>
<tr>
<th>Group 1 (THY)</th>
<th>Group 2 (GRA)</th>
<th>Group 3 (MAR)</th>
<th>Group 4 (LAM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.77%</td>
<td>37.62%</td>
<td>28.06%</td>
<td>27.25%</td>
</tr>
</tbody>
</table>

Table 1: Percentage of missing values per group of samples

We decided whether the missing values within the data are censored or MCAR by inspecting the distributions of the mean values of the non-missing values per proteins. MVs associated to proteins for which the mean value of the non-missing values is below a threshold are considered as being censored, while the remaining MVs are considered as MCAR.

The histograms of the mean value of non-missing values for all proteins and for each group of samples are shown in Figure 1 below:

![Figure 1: Histogram of the mean values of the log-transformed non-missing values per protein for each group of samples (lamellae, margins, grana).](image)

From Figure 1 one can easily observe that most of the MVs are recorded for proteins which exhibit very low abundance in the samples where they have been detected. Therefore the MVs imputation strategy is based on an empirical threshold; MV corresponding to proteins for which the mean value...
of the non-MV is less than 2 are considered as being censored and they have been imputed by the minimum value observed, while the remaining MV are considered as being missing at random, and they have been imputed using a nearest neighbor strategy (Troyanskaya, 2001).

**Discussion:** One could argue about the choice of the empirical threshold used here. This empirical threshold corresponds to a decision boundary of declaring missing values as being censored or MCAR. The distributions depicted in Figure 1 show long tails and prominent peaks when the mean value of non-missing values is close to 1, meaning that most of the proteins present censored MVs. Proteins with mean values of the non-MVs found in the tail of the distributions shown above are considered as being outliers of a “normal behavior” of the mean values associated to proteins presenting censored MVs; therefore, we considered them as being MCAR, and the decision threshold has been empirically chosen by inspecting the distributions in Figure 1.

**Data normalization**

For the 24 samples corresponding to the 4 biological groups (thylakoids, grana-BBY, margins and lamellae), the distributions of the log-transformed abundances are examined in order to investigate about the presence of technical bias, see Figures 2. Note that the sample normalization has been performed after MVs imputation, which explains the long-tailed distributions with a prominent peak in 0 for the grana-BBY, margins and lamellae samples which present high rates of missing proteins. Quintiles normalization has been performed within each group of samples in order to render the sample distributions homogeneous. The sample-wise boxplots for each group are shown in Figure 3.

![Boxplots](image.png)

**Figure 2:** Sample-wise boxplots of log-transformed abundances before normalization (with missing values).
values imputed - note that especially for Grana-BBY, Margins and Lamellae which present high rates of missing proteins and missing values, the distributions present a prominent peak close to 0, because most of the missing values are censored, hence imputed with the minimum value observed).

**Figure 3:** Sample-wise boxplots of log-transformed abundances after normalization (with missing values imputed).

**Tuning the number of clusters**

The decision upon the number of clusters was taken by analyzing both the goodness of fit as well as the stability of a wide range of clustering solutions. By doing so, we ensure that the chosen solution on one hand, will find the data structures which best fit the actual data in terms of intergroup separability and cohesion, and on another hand, will reveal the most stable data structures; this means that the chosen solution will reveal distinct groups of similar proteins and at the same time, ensuring that those groups are globally the most stable among the solutions investigated. As there is no golden standard for assessing the goodness of fit of a clustering solution, this was assessed by investigating several indices for clustering solutions ranging from 2 to 20. In our analysis we selected 4 indices that have been highly ranked in a comparative study evaluating a number of 30 such goodness of fit procedures (1): Calinski-Harabasz index (CH) (2), Gamma index (Gamma) (3), silhouette coefficient (4) and log(SSB/SSW) (5). Complementary to cluster validation by goodness of fit indices, we also investigated the stability of clustering solutions as for the goodness of fit analysis (solutions ranging between 2 and 20 clusters). Cluster stability basically accounts for the sensibility of a clustering solution against perturbation of the data and it is a well-settled strategy for cluster
Thylakoid grana and stroma-lamellae proteomes validation. The strategy used consisted of the following steps (6): (i) Run the clustering method on the original dataset, (ii) Resample by bootstrapping new datasets from the original one and apply the clustering method on them (other strategies for stability testing employed here consist in randomly replacing samples with noise samples as well as in adding jitter to the data), (iii) For every cluster obtained from the original data find the most similar cluster in the new clustering and record the similarity value which is the Jaccard coefficient between two subsets of a set, based on set memberships, (iv) Assess the cluster stability of every single cluster by the mean Jaccard coefficient taken on the resample datasets. Finally, we assessed the overall cluster stability of a cluster solution by the mean value of the Jaccard coefficients taken on the all clusters. For cluster solutions ranging from 2 to 20, these values are shown in Figure 4. Cluster analysis was performed using the R package fpc (6). The cluster solution with 7 clusters has been decided as a trade-off between the goodness of fit and the stability analysis. Several solutions can be considered by analyzing the goodness of fit indices: the CH index is rather uninformative, as it exhibits many local maxima, the global maxima being found when the number of clusters equals 12, which appears to be rather an unstable solution. The Gamma index shows a global maximum for a number of clusters equal to 4, but solutions with 6 and 7 clusters also exhibit relatively high goodness of fit values. The silhouette coefficient has its global maximum when the number of clusters equals 7, while according to the log(SSW/SSB) index a solution with 8 clusters could be considered. The stability indices exhibit their maximum values when the number of clusters equals 3 respectively 4 clusters, however, these solutions have low values for the goodness of fit coefficients. Solutions with 7 clusters also present high stability values while being in the same time well-fitted to the data. As a 7 clusters solution is most indicated by the goodness of fit indices, we decided to cluster the 1040 proteins identified in the thylakoid fraction in 7 groups.
Thylakoid grana and stroma-lamellae proteomes

<table>
<thead>
<tr>
<th>c)</th>
<th>d)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Stability boot" /></td>
<td><img src="image" alt="Stability noise" /></td>
</tr>
<tr>
<td><img src="image" alt="Stability jitter" /></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4. Validation of clustering solutions: goodness of fit indices a) – Calinsky-Harabasz index, b) – Gamma index c) – silhouette coefficients and d) – log-ratio within-between sum of squares; stability of clustering solutions when data is d) resampled by bootstrapping, e) replaced by noise, f) jittered. For indices a), b), c), e), f) and g) global or high local maxima could indicate good clustering solutions. According to the log-ratio within-between sum of squares, the presence of “knees” in the outlined plot should reveal good clustering solutions.
References


