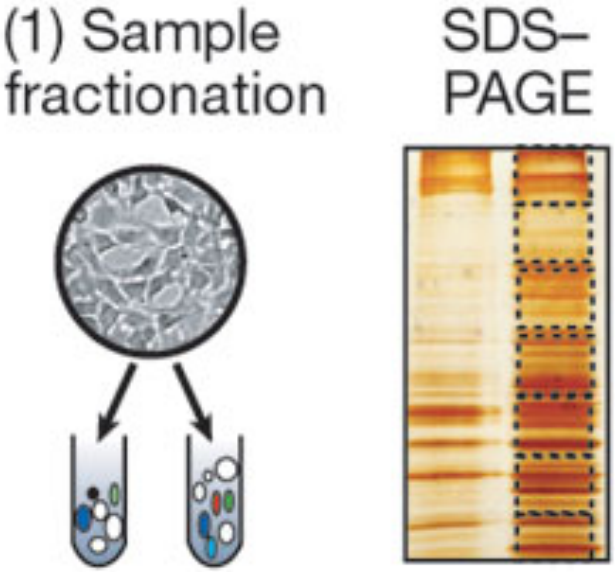
V3 – MS proteomics – data imputation

- How does MS proteomics work?
- What is the role of bioinformatics in MS proteomics ?
  - Peptide mass fingerprinting
  - Significance analysis
  - GO annotations
- Applications of MS:
  - TAP-MS
  - Phosphoproteome
- **Data imputation** for MS data
  - Identify TRAP clients

Noble prize in chemistry 2002
John B. Fenn Koichi Tanaka
"for their development of soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules"

www.nobelprize.org

Proteomics workflow: (1) protein isolation



The typical proteomics experiment consists of 5 stages.

In stage 1, the proteins to be analyzed are **isolated** from cell lysate or tissues by biochemical fractionation or affinity selection.

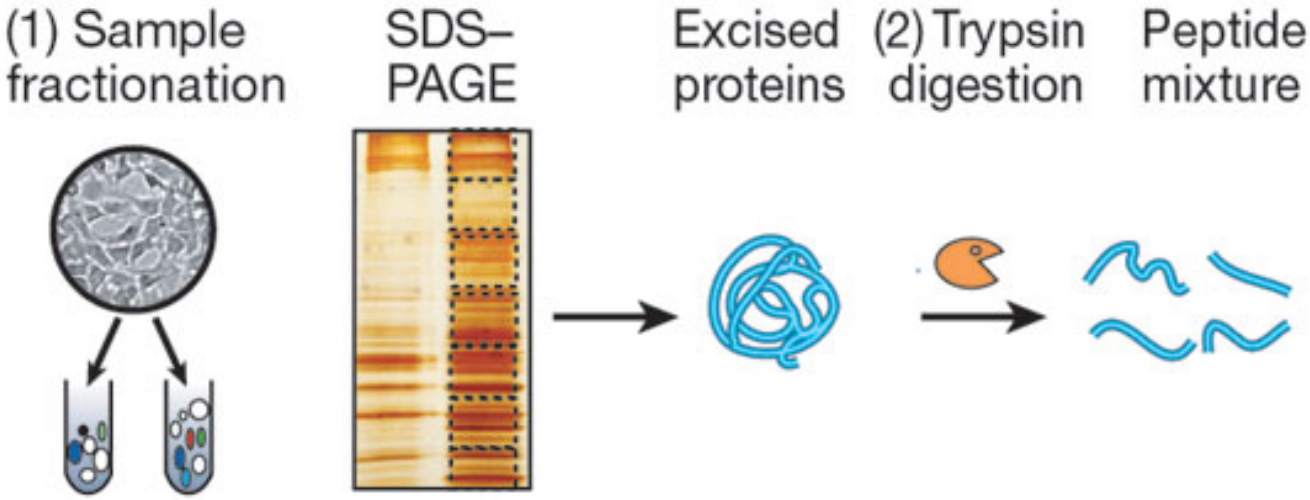
This often includes a final step of one-dimensional gel electrophoresis, and defines the 'sub-proteome' to be analysed.

MS of whole proteins is less sensitive than **peptide MS.**

The mass of the intact protein by itself is insufficient for identification.

Aebersold, Mann
Nature 422, 198-207(2003)

Proteomics workflow: (2) trypsin digestion



Therefore, in stage 2, proteins are **degraded enzymatically** to peptides, usually by trypsin.

This yields peptides with C-terminally protonated amino acids (K/R) which is beneficial in subsequent peptide sequencing.

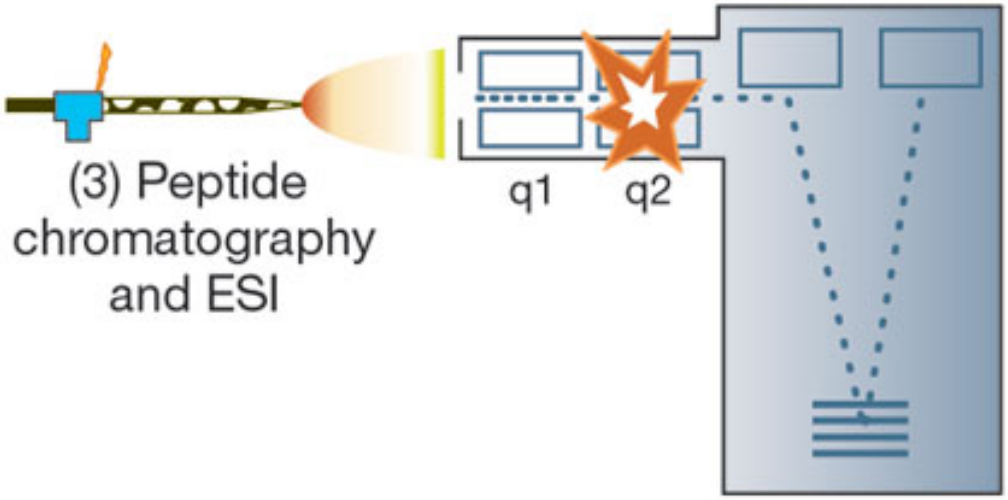
Table 1. Distrubution of peptide fragment length from 20,639 proteins

Enzyme/reagent	Residues cleaved	Total fragments	Avg. fragment length
Trypsin	K/R	662,981	8
Lys-C	K	359,140	16
Asp-N	D	321,655	18
CNBr	M	150,605	38
Hydroxylamine	N-G	36,643	152
Dilute acid	D-P	35,574	166

Aebersold, Mann
Nature 422, 198-207(2003)

Henzel et al. J Am Soc Mass Spectrom
14, 931–942 (2003)

Proteomics workflow: (3) peptide chromatography



In stage 3, the peptides are **separated** by one or more steps of high-pressure liquid chromatography in very fine capillaries.

Then, they are eluted e.g. into an electrospray ion source where they are **nebulized** in small, **highly charged droplets**.

After evaporation, multiply protonated peptides enter the mass spectrometer.

Aebersold, Mann
Nature 422, 198-207(2003)

Mass spectrometer

A mass spectrometer consists of an **ion source**, a **mass analyser** that measures the **mass-to-charge ratio** (*m*/*z*) of the ionized analytes, and a **detector** that registers the number of ions at each *m*/*z* value.

Electrospray ionization (ESI) and **matrix-assisted laser desorption/ionization** (MALDI) are the two techniques most commonly used to volatize and ionize the proteins or peptides for mass MS analysis.

ESI ionizes the analytes out of a **solution** and is therefore readily coupled to liquid-based (e.g. chromatographic and electrophoretic) separation tools.

MALDI sublimates and ionizes the samples out of a **dry, crystalline matrix** via laser pulses.

MALDI-MS is normally used to analyse relatively simple peptide mixtures, whereas integrated liquid-chromatography ESI-MS systems (LC-MS) are preferred for the analysis of complex samples

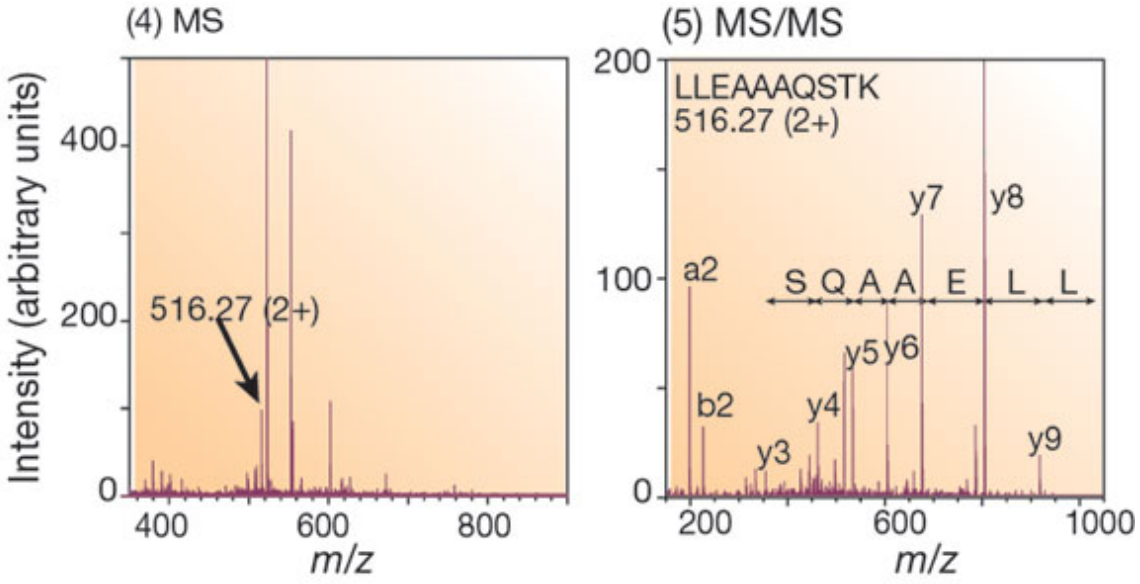
Aebersold, Mann
Nature 422, 198-207(2003)

Proteomics workflow: (4) MS

In stage 4, a mass spectrum of the peptides eluting at this time point is taken.

Mass peak ≡ **sequence composition** of a peptide.

The computer then generates a prioritized list of the peptides for a second fragmentation.



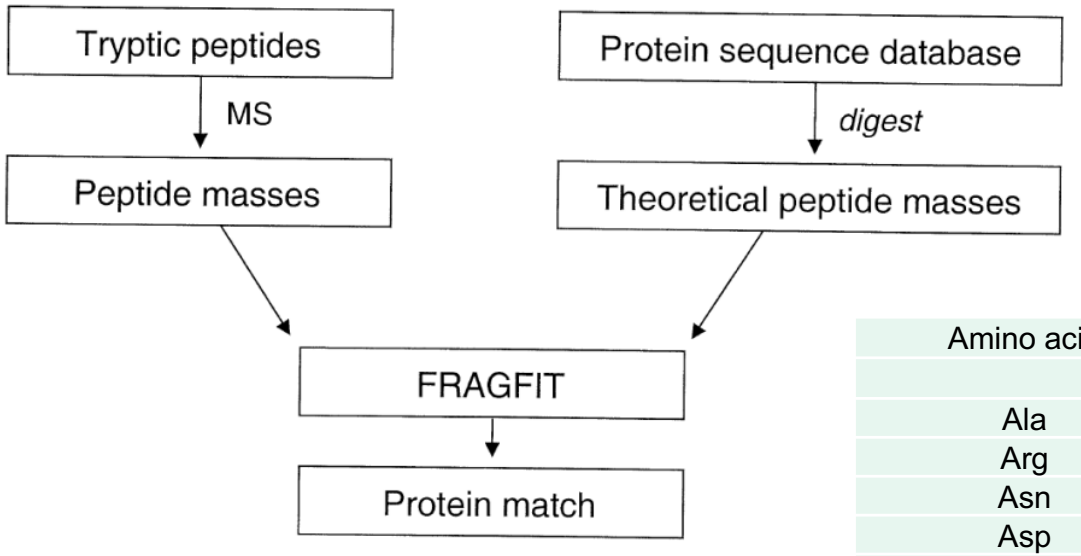
In stage 5, a series of **tandem mass spectrometric** or 'MS/MS' experiments is performed to determine the sequence of a peptide (here, the peak *m* = 516.27 Da).

The MS and MS/MS spectra are matched against protein sequence databases ("**peptide mass fingerprinting**").

The outcome of the experiment is the identity of the peptides and therefore the proteins making up the purified protein population.

Aebersold, Mann
Nature 422, 198-207(2003)

Peptide mass fingerprinting



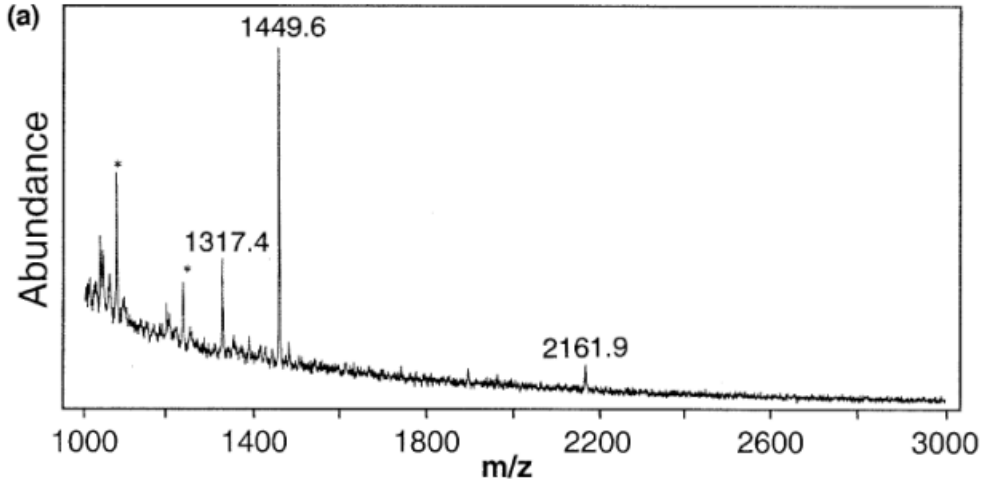
$$m_{peptide} = \sum_{i \in amino\ acids\ 1 \ldots n} m_i$$

Amino acid	Mono-Isotopic mass [Da]	Average mass [Da]
Ala	71.037114	71.0779
Arg	156.101111	156.1857
Asn	114.042927	114.1026
Asp	115.026943	115.0874
Cys	103.009185	103.1429
Glu	129.042593	129.114
Gln	128.058578	128.1292
Gly	57.021464	57.0513
His	137.058912	137.1393
Ile	113.084064	113.1576
Leu	113.084064	113.1576
Lys	128.094963	128.1723
Met	131.040485	131.1961
Phe	147.068414	147.1739
Pro	97.052764	97.1152
Ser	87.032028	87.0773
Thr	101.047679	101.1039
Trp	186.079313	186.2099
Tyr	163.06332	163.1733
Val	99.068414	99.1311

The masses of peptides from a database are compared with experimentally determined masses using a software.

Henzel et al. J Am Soc Mass Spectrom 14, 931–942 (2003);
www.matrixscience.com

Peptide mass fingerprinting



(b)
```
enzyme: Asp-N (N-side of Asp)
Mass of MH+: 1317.400 1449.600 2161.900 (tol: 1.000)
LZCH Lysozyme c (EC 3.2.1.17) precursor - Chicken
          2162.444          84: DGRTPGSRNLCNIPCSALLSS
          1449.706         105: DITASVNCAKIVS
          1317.552         137: DVQAWIRGCRL
```

Mass [Da]

Starting position

Peptide fragment

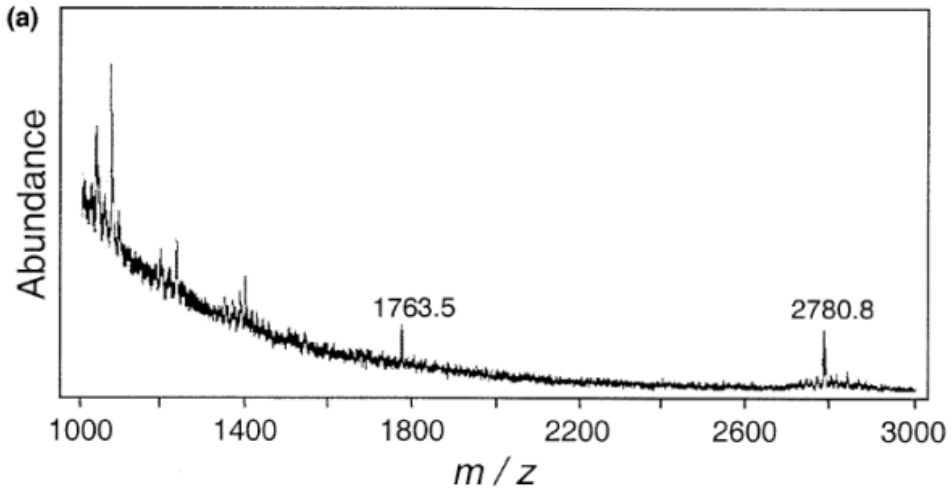
(**a**) FAB ("fast atom bombardment", an old technique) spectrum of a 250 pmol tryptic digest of Asp-N digest of **lysozyme**.

3 characteristic peaks are labeled.

(**b**) FRAGFIT output page showing a match with chicken egg white lysozyme obtained using the masses from the MS spectrum.

Henzel et al. J Am Soc Mass Spectrom 14, 931–942 (2003)

Peptide mass fingerprinting



(b)

```
enzyme: CNBr (C-side of Met)
Mass of MH+: 1763.500 2780.800 (tol: 0.600)
CCHO Cytochrome C - Horse
          1764.031          66: EYLENPKKYIPGTKM
          2781.268          81: IFAGIKKKTEREDLIAYLKKATNE
CCHOD Cytochrome C - Donkey and common zebra
                     (tentative sequences)
          1764.031          66: EYLENPKKYIPGTKM
          2781.268          81: IFAGIKKKTEREDLIAYLKKATNE
```

(**a**) FAB spectrum of a 500 pmol CNBr cleavage of horse heart **cytochrome *c***.

(**b**) FRAGFIT output page showing a match with cytochrome *c* obtained using the masses from the FAB spectrum.

The output includes all proteins that match the mass list.

The 2 masses observed were sufficient to identify the protein as cytochrome c and permitted the identification of the species.

At the time this search was performed, the database contained nearly 100 different species of cytochrome *c*

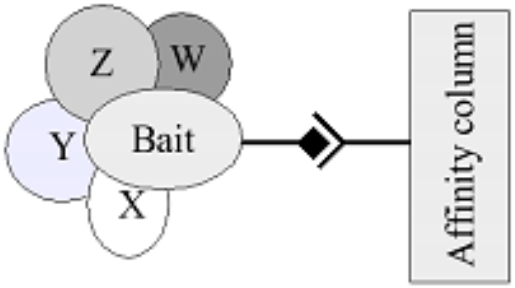
Henzel et al. J Am Soc Mass Spectrom 14, 931–942 (2003)

Application: Detect protein-protein interactions: Tandem affinity purification (also „pull-down“)

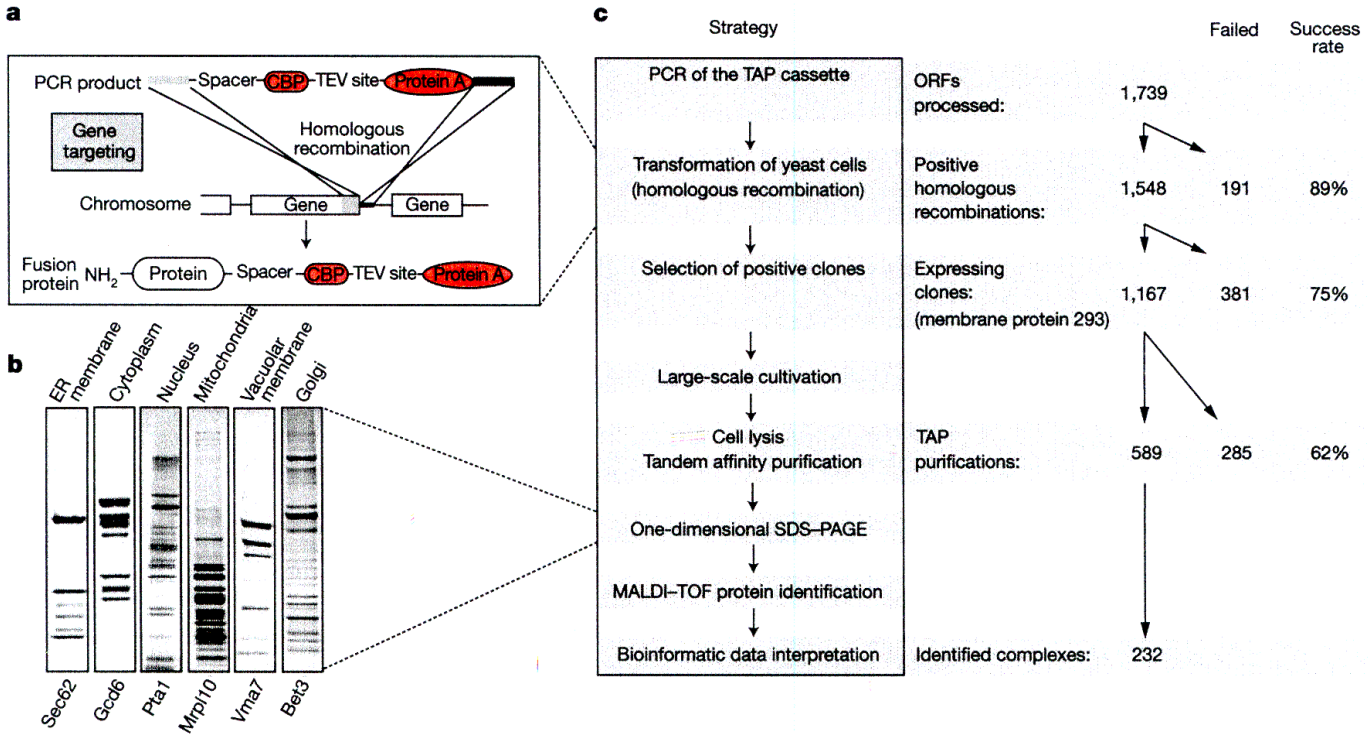
In **affinity purification**, a protein of interest (bait) is tagged with a molecular label (dark route in the middle of the figure) to allow easy purification.

The tagged protein is then co-purified together with its interacting partners (W–Z).

This strategy can be applied on a genome scale (as Y2H).



Identify proteins by mass spectro-metry (MALDI-TOF).

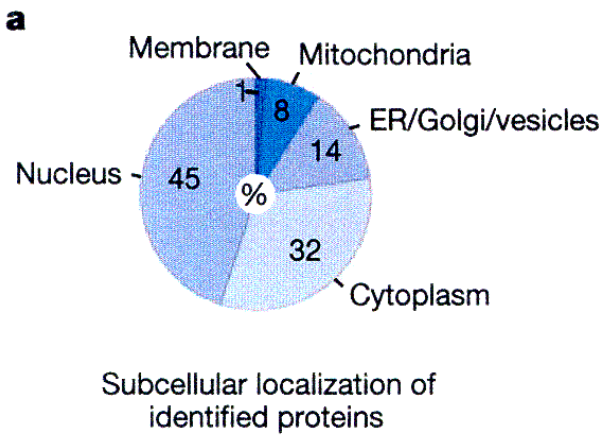


Gavin *et al. Nature* 415, 141 (2002)

TAP analysis of yeast PP complexes

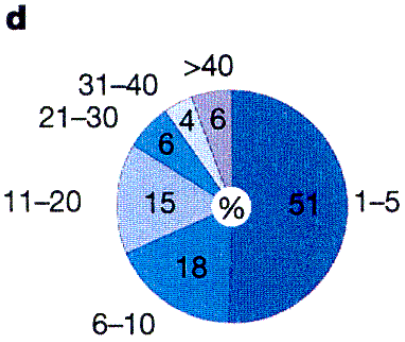
Identify proteins by scanning yeast protein database for protein composed of fragments of suitable mass.

(a) lists the identified proteins according to their localization
-> no apparent bias for one compartment, but very few membrane proteins (should be ca. 25%)



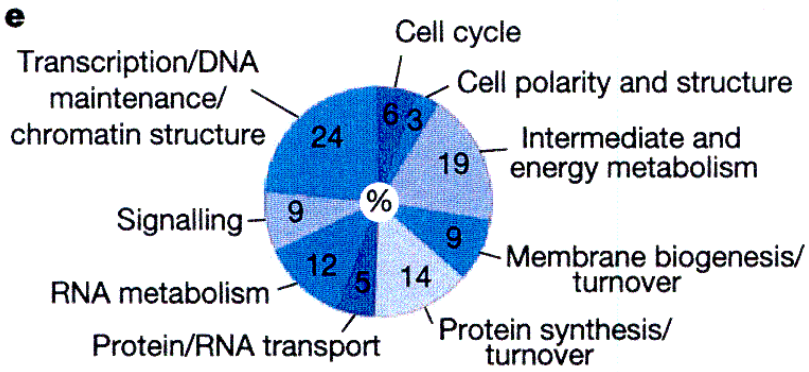
Subcellular localization of identified proteins

(d) lists the number of proteins per complex
-> half of all PP complexes have 1-5 members, the other half is larger



Number of proteins per complex

(e) Complexes are involved in practically all cellular processes



Distribution of complexes according to function

Gavin *et al. Nature* 415, 141 (2002)

Application of MS: Protein phosphorylation during cell cycle

Protein **phosphorylation** and **dephosphorylation** are highly controlled biochemical processes that respond to various intracellular and extracellular stimuli.

Phosphorylation status modulates **protein activity** by:
- influencing the tertiary and quaternary **structure** of a protein,
- controlling **subcellular distribution**, and
- regulating its **interactions** with other proteins.

Regulatory protein phosphorylation is a **transient** modification that is often of low occupancy or “stoichiometry”

This means that only a fraction of a particular protein may be phosphorylated on a given site at any particular time, and that occurs on regulatory proteins of low abundance, such as protein kinases and transcription factors.

Olsen Science Signaling 3 (2010)

Cell Cycle and the Phosphoproteome

CELL CYCLE

Quantitative Phosphoproteomics Reveals Widespread Full Phosphorylation Site Occupancy During Mitosis

Jesper V. Olsen,[1,2*] Michiel Vermeulen,[1,3*] Anna Santamaria,[4*] Chanchal Kumar,[1,5*] Martin L. Miller,[2,6] Lars J. Jensen,[2] Florian Gnad,[1] Jürgen Cox,[1] Thomas S. Jensen,[7] Erich A. Nigg,[4] Søren Brunak,[2,7] Matthias Mann[1,2†]

(Published 12 January 2010; Volume 3 Issue 104 ra3)

www.SCIENCESIGNALING.org 12 January 2010 Vol 3 Issue 104 ra3

Aim: Analyze all proteins that are modified by phosphorylation during different stages of the cell cycle of human HeLa cells.

Ion-exchange chromatography + HPLC + MS + sequencing led to the identification of 6695 proteins.
From this 6027 quantitative cell cycle profiles were obtained.

A total of 24,714 phosphorylation events were identified.
20,443 of them were assigned to a specific residue with high confidence.

Finding: about **70%** of all human proteins get phosphorylated.

Review: protein quantification by SILAC

ARTICLE

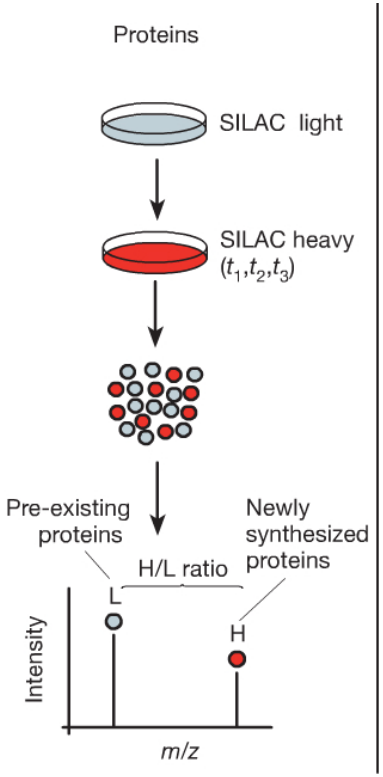
doi:10.1038/nature10098

Global quantification of mammalian gene expression control

Björn Schwanhäusser[1], Dorothea Busse[1], Na Li[1], Gunnar Dittmar[1], Johannes Schuchhardt[2], Jana Wolf[1], Wei Chen[1] & Matthias Selbach[1]

SILAC: „**s**table **i**sotope **l**abelling by **a**mino acids in **c**ell culture“ means that cells are cultivated in a medium containing heavy stable-isotope versions of essential amino acids.

When non-labelled (i.e. light) cells are transferred to heavy SILAC growth medium, newly synthesized proteins incorporate the heavy label while pre-existing proteins remain in the light form.



Quantification protein turnover and levels.
Mouse fibroblasts are transferred to medium with heavy amino acids (SILAC).

Protein turnover is quantified by mass spectrometry and next-generation sequencing, respectively.

Schwanhäuser et al. Nature 473, 337 (2011)

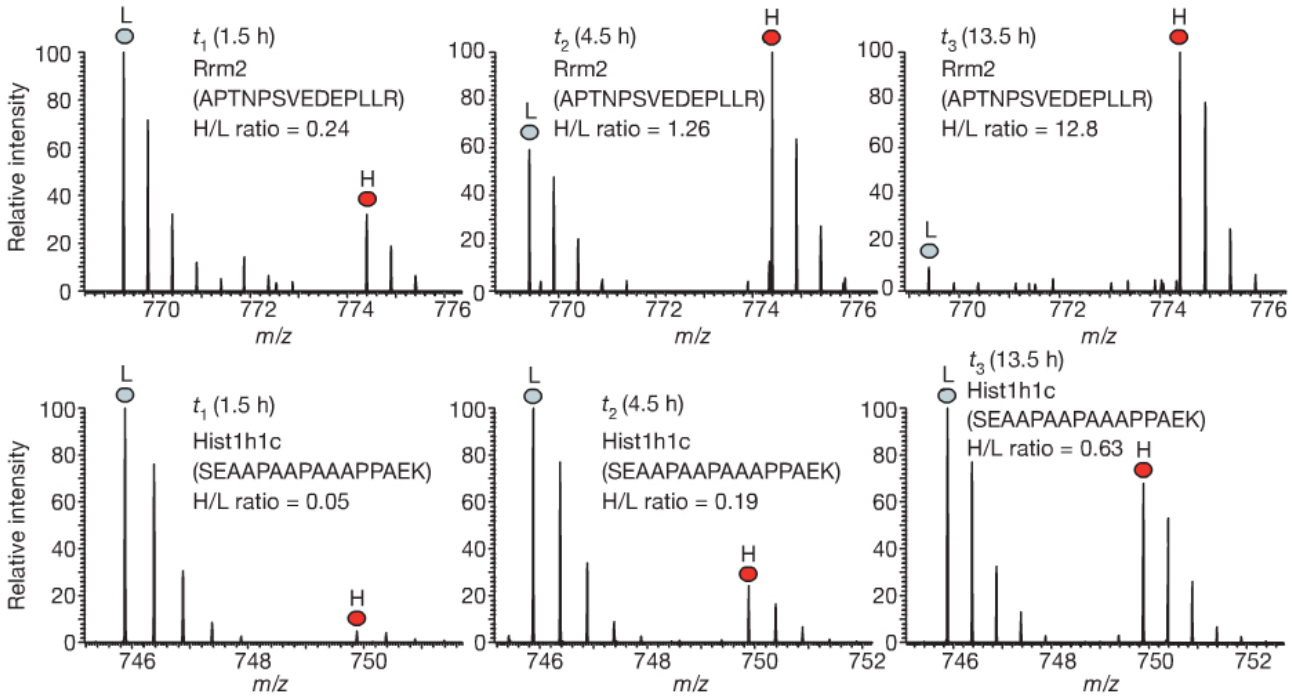
Rates of protein translation

Mass spectra of peptides for two proteins.

Top: **high-turnover protein**
Bottom: **low-turnover protein**.

Over time, the heavy to light (H/L) ratios increase.

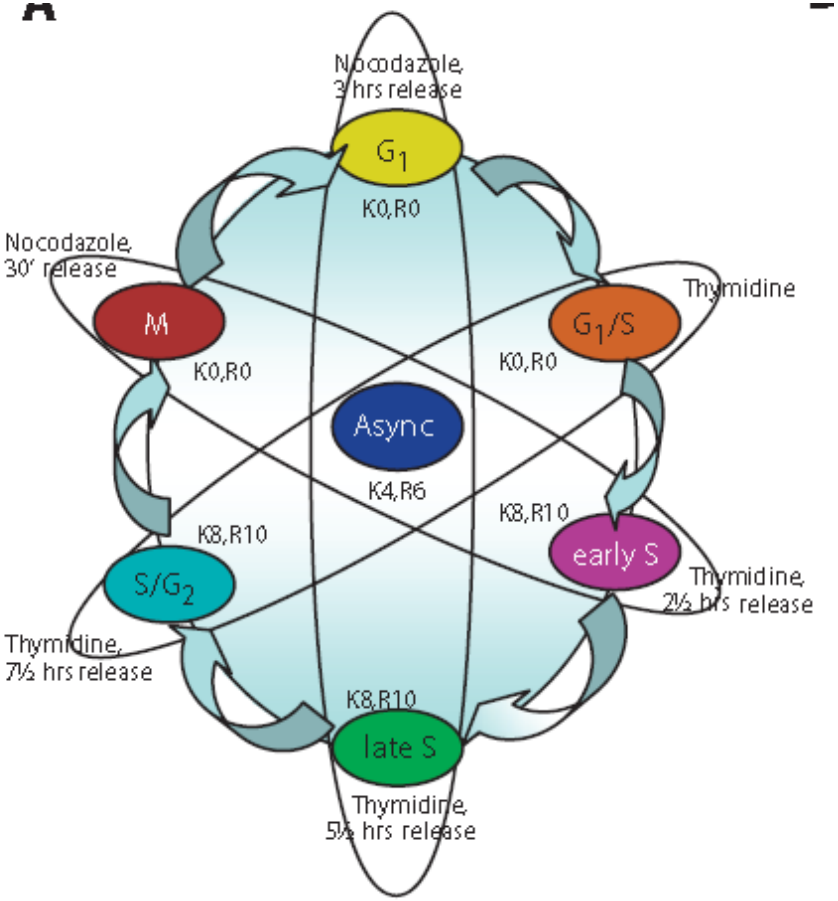
H-concentration of high-turnover protein saturates.
That of low-turnover protein still increases.



This example was introduced to illustrate the principles of SILAC and mass spectroscopy signals (peaks).
In the Olson et al. study, the authors used H and L forms to label different stages of the cell cycle.

Schwanhäuser et al. Nature 473, 337 (2011)

Quantitative proteomic analysis



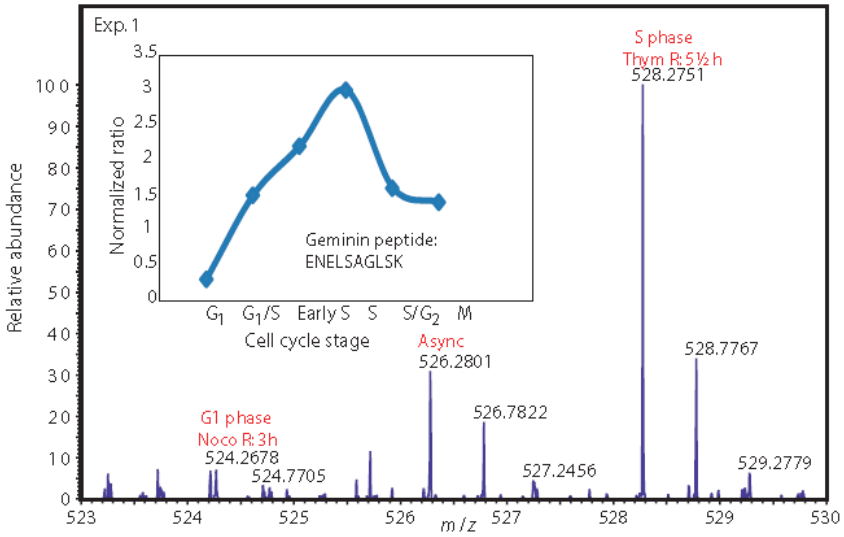
- HeLa S3 cells were SILAC-labeled with 3 different isotopic forms (light – medium –heavy) of arginine and lysine.

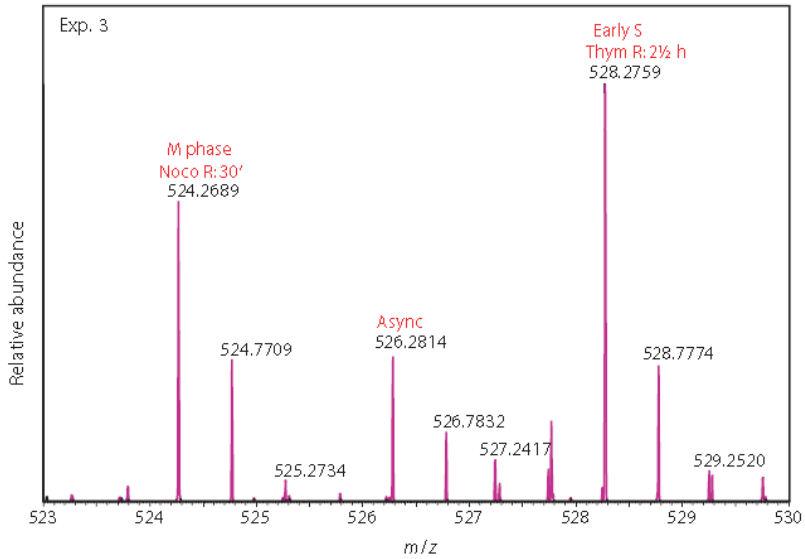
3 individual populations of heavy and light SILAC cells were synchronized with a **thymidine** block (analog of thymine, blocks entry into S phase). Cells were then collected at 6 different time points across the cell cycle after release from the thymidine arrest.

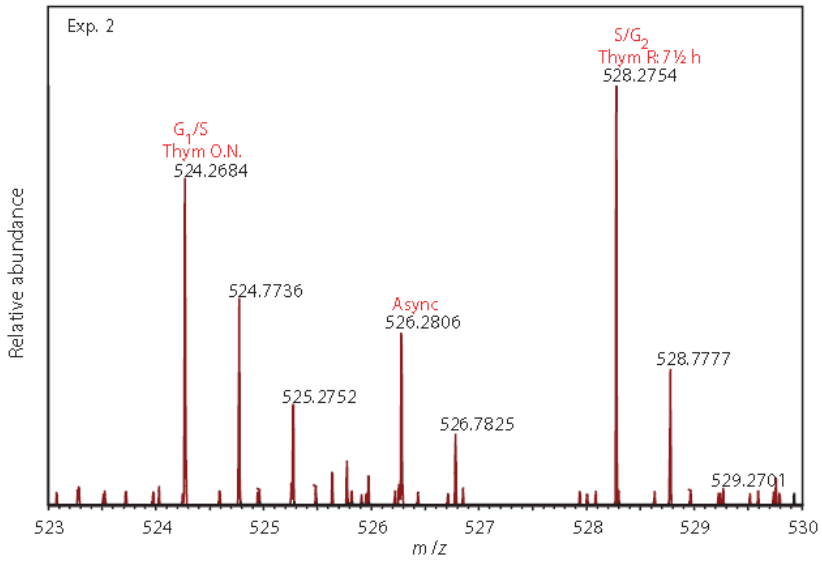
2 samples were collected after a **cell cycle arrest** with **nocodazole** and release. (Nocodazole interferes with polymerization of microtubules.)

Cells were lysed and mixed in equal amounts using an asynchronously growing cell population as the internal standard to allow normalization between experiments.
3 independent experiments were performed to cover six cell cycle stages.

Monitoring of protein abundance by MS







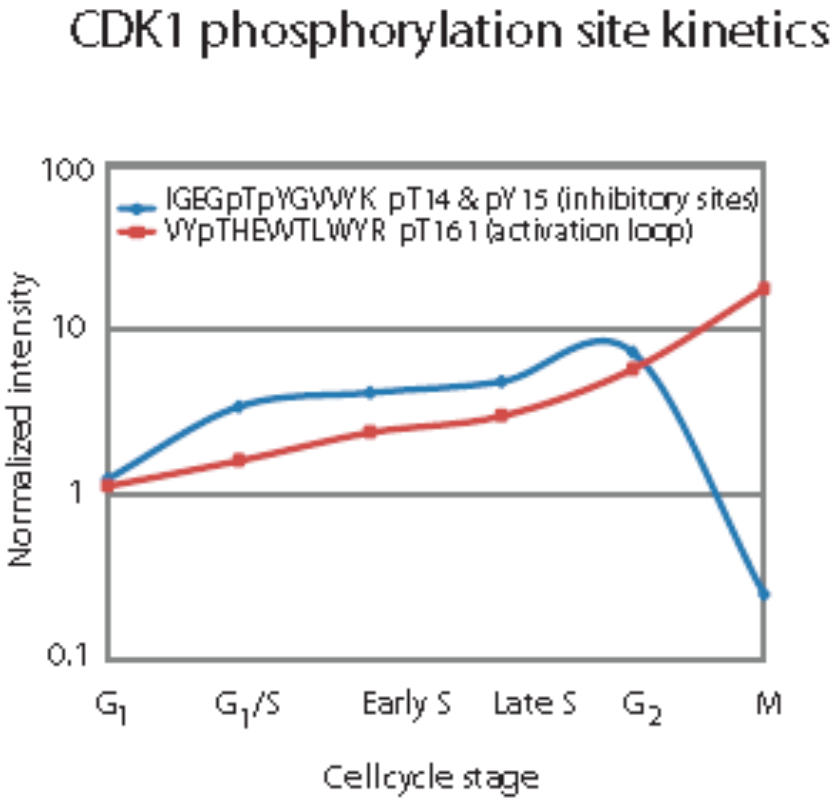
Representative MS data showing how the abundance of the proteins was monitored in three experiments (Exp. 1, Exp. 2, Exp. 3) to obtain information from the 6 stages of the cell cycle.

The data show the MS analysis of a tryptic SILAC peptide triplet derived from the cell cycle marker protein **Geminin**.

Relative peptide abundance changes were **normalized** to the medium SILAC peptide derived from the **asynchronously** grown cells in all three experiments.

Inset: combined six-time point profile of Geminin over the cell cycle.

Example: Dynamic phosphorylation of CDK1

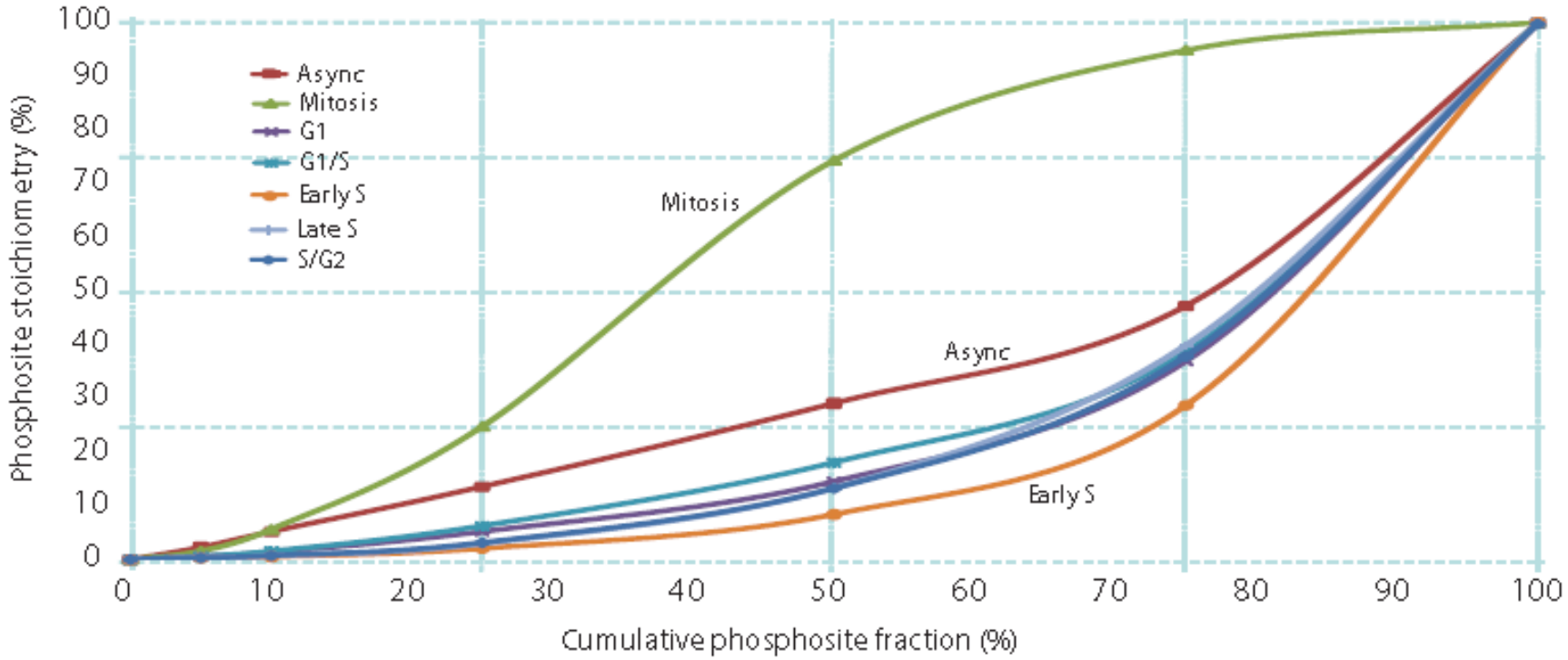


Dynamic profile of two CDK1 phosphopeptides during the cell cycle.

The activating site T161 (**red**) peaks in mitosis, whereas phosphorylation of the inhibitory sites T14 and Y15 (**blue**) is decreased in mitosis

Olsen Science Signaling 3 (2010)

Total phosphosite occupancy in different stages of cell cycle



Fifty percent of all mitotic phosphorylation sites have occupancy of 75% or more.

Olsen Science Signaling 3 (2010)

Data imputation

What is the role of data imputation in MS data?

If no signal is detected, this can have various reasons:

- The peptide is not detected or falsely identified

- The peptide is really not at all present in the sample

- The peptide concentration is below the detection threshold …

The reason for missing data is generally not known.

Simply setting all missing data to zero would generate **false positive** signals
= proteins appear to be significantly deregulated, but are in fact not.

Imputation methods: KNNimpute

Lets assume that gene $\mathbf{g_1}$ lacks data point *i* and the total number of genes is *m*.

The KNNimpute method (Troyanskaya *et al.*, 2001) finds *k* (*k* < *m*) other genes with expressions most similar to that of $\mathbf{g_1}$ and that do have a measured value in position *i*.

The missing value of $\mathbf{g_1}$ is estimated by the weighted average of the values in position *i* of these *k* closest genes.

$$\mathbf{g}^* = \frac{\omega_1 \mathbf{g}_{s_1} + \omega_2 \mathbf{g}_{s_2} + \cdots + \omega_k \mathbf{g}_{s_k}}{\omega_1 + \cdots + \omega_k},$$

Here, the contribution of each gene is weighted by the similarity of its expression to that of $\mathbf{g_1}$.

Kim et al., Bioinformatics 21, 187 (2005)

Imputation methods: SVDimpute

SVDimpute method (Troyanskaya *et al.*, 2001):

- Given: matrix G where some data is missing.
- Generate initial matrix *G'* from G by substituting all missing values of the *G* by zero or row averages.
- Compute SVD of G'.
- Determine the *t* most significant eigengenes of *G'* (with largest eigenvalues).
- Regress every gene with missing values against the *t* most significant eigengenes (by ignoring position *i*)

Using the coefficients of the regression, the missing value in G is estimated as a linear combination of the values in the respective position *i* of the *t* eigengenes.

This procedure is repeated until the total change of the matrix G' becomes insignificant.

Kim et al., Bioinformatics 21, 187 (2005)

Imputation methods: Local Least squares

(1) select *k* genes that have similar properties (e.g. expression profiles) to the gene where position *i* is missing.

Similarity can be based on the *L2*-norm or Pearson correlation coefficients of the expression profiles.

(2) regression and estimation

Kim et al., Bioinformatics 21, 187 (2005)

Imputation methods: Local Least squares

Based on the k neighboring gene vectors, form the matrix $A \in \mathrm{R}^{k \times (n-1)}$ and the two vectors $\mathbf{b} \in \mathrm{R}^{k \times 1}$ and $\mathbf{w} \in \mathrm{R}^{(n-1) \times 1}$.

The k rows of the matrix A consist of the k-nearest neighbor genes $\mathbf{g}^{\mathrm{T}}_i \in \mathrm{R}^{1 \times n}$, $1 \le i \le k$, with position i deleted.

The elements of the vector $\mathbf{b}$ consists of position i of the k vectors $\mathbf{g}^{\mathrm{T}}_i$.
The elements of the vector $\mathbf{w}$ are the $n - 1$ elements of the gene vector $\mathbf{g}_1$ whose missing position i is deleted.

After the matrix A, and the vectors $\mathbf{b}$ and $\mathbf{w}$ are formed, the least squares problem is formulated as

$$\min_{\mathbf{x}} \| A^{\mathrm{T}}\mathbf{x} - \mathbf{w} \|_2$$

Then, the missing value α is estimated as a linear combination of the respective values of the neighboring genes

$$\alpha = \mathbf{b}^{\mathrm{T}}\mathbf{x} = \mathbf{b}^{\mathrm{T}}(A^{\mathrm{T}})^{\dagger}\mathbf{w}$$

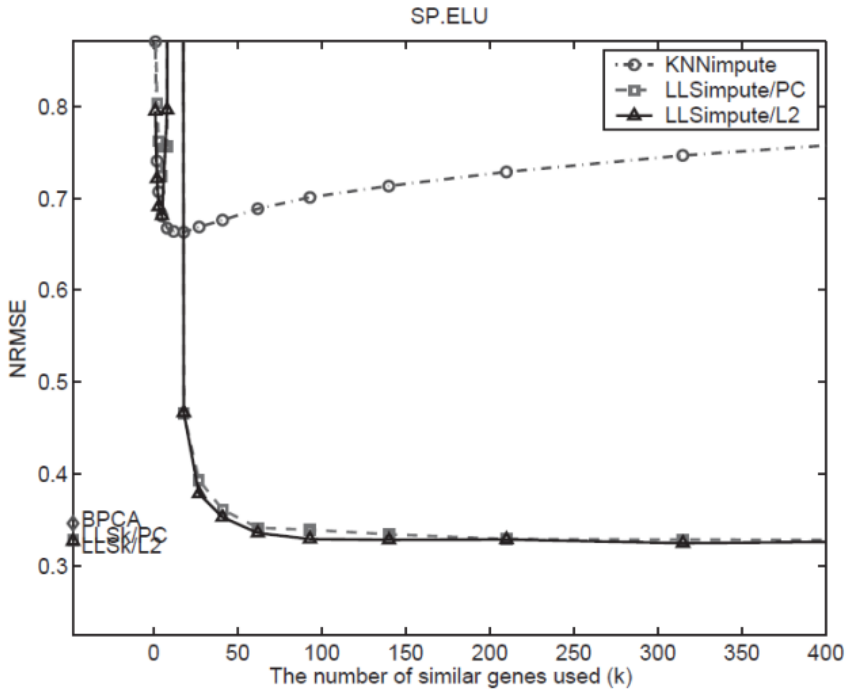
Kim et al., Bioinformatics 21, 187 (2005)

Imputation methods: Local Least squares

Spellman data set: yeast cell cycle
5% of data were missing

-> LLSimpute outperforms KNNimpute

Lower Root Mean Square Error (RMSE)



Kim et al., Bioinformatics 21, 187 (2005)

Models for missing values

Missing Completely At Random (MCAR): in a proteomics data set, this corresponds to the combination of a propagation of multiple minor errors or stochastic fluctuations. e.g. by a misidentified peptide

Missing At Random (MAR): this is a more general class than MCAR, where conditional dependencies are accounted for. In a proteomics data set, it is classically assumed that all MAR values are also MCAR.

Missing Not At Random (MNAR) assumes a **targeted effect**. E.g. in MS-based analysis, chemical species whose abundances are close enough to the limit of detection of the instrument record a higher rate of missing values.

Imputation methods for MCAR and MAR are general.
For MNAR, they are methods-specific.

Lazar et al., J Proteome Res 15, 1116 (2016)

Simulation benchmark

Use real data (Super-SILAC and label-free quantification) on human primary tumor-derived xenograph proteomes for the two major histological subtypes of nonsmall cell lung cancer : adenocarcinoma and squamous cell carcinoma.

MNAR values: one randomly generates a **threshold matrix** T from a Gaussian distribution with parameters (μ_t = q, σ_t = 0.01), where q is the α-th quantile of the abundance distribution in the complete quantitative data set.
Then, each cell (*i,j*) of the complete quantitative data set is compared with $T_{i,j}$.
If $(i,j) \geq T_{i,j}$, the abundance is not censored.
If $(i,j) < T_{i,j}$, a Bernoulli draw with probability of success βα · 100 determines if the abundance value is censored (success) or not (failure).
α and β are the rate of missing values and the MNAR ratio, respectively.

MCAR values are incorporated by replacing with a missing value the abundance value of n m ((100 - β) α /100) randomly chosen cells in the table of the quantitative data set.

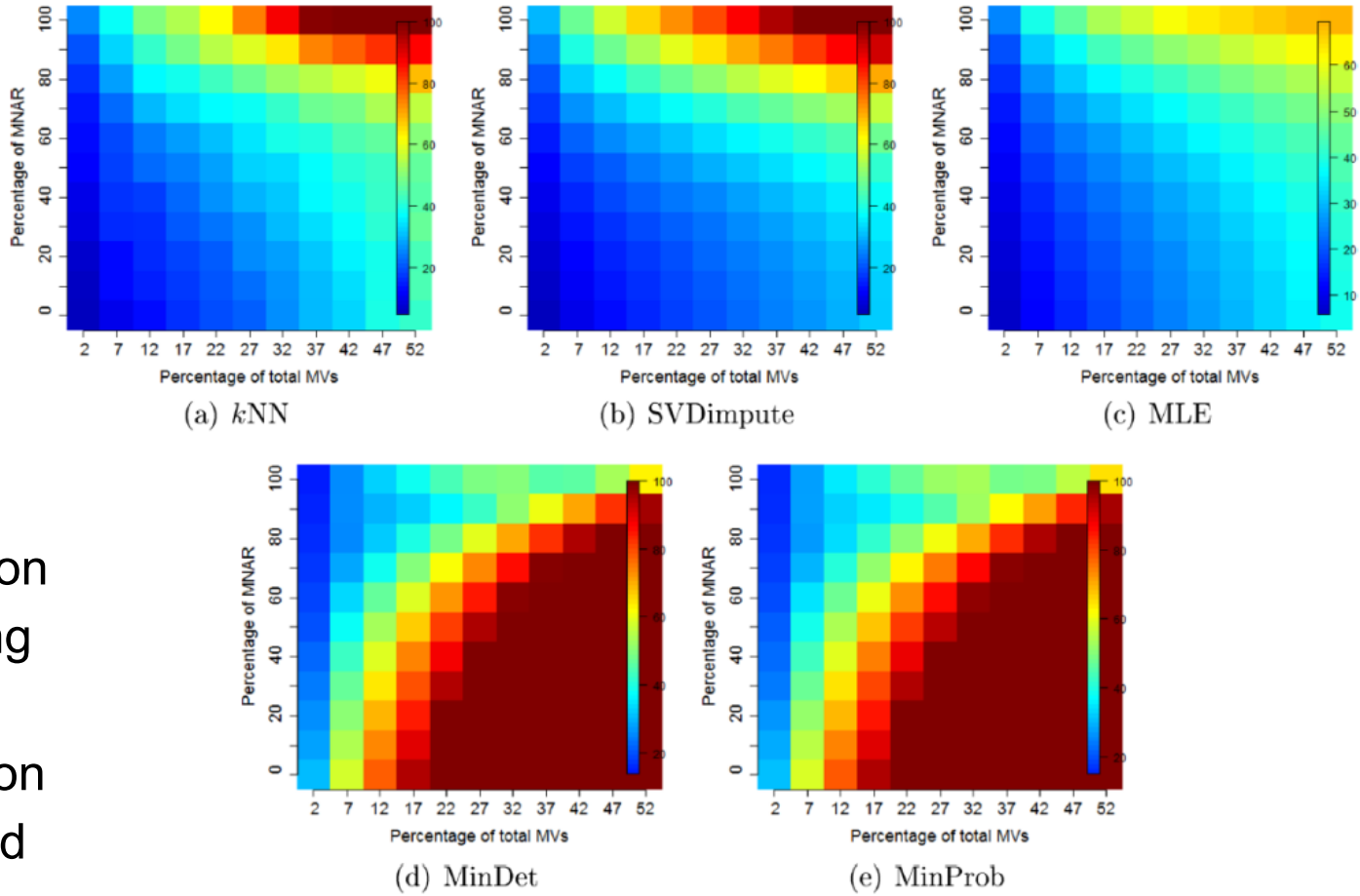
Lazar et al., J Proteome Res 15, 1116 (2016)

Imputation methods: benchmark

MLE: maximum likelihood estimator

MinDet: simply replace missing values by the minimum value that is observed in the data set.

MinProb: stochastic version of MinDet. Replace missing values with random draws from a Gaussian distribution centered on the value used with MinDet and with a variance tuned to the median of the peptide-wise estimated variances



RSR = RMSE / std.dev. MV: missing value

Blue: low RSR

Red: high RSR

Lazar et al., J Proteome Res 15, 1116 (2016)

Conclusion on data imputation

Algorithms SVDimpute, kNN, and MLE perform better under a small MNAR ratio.

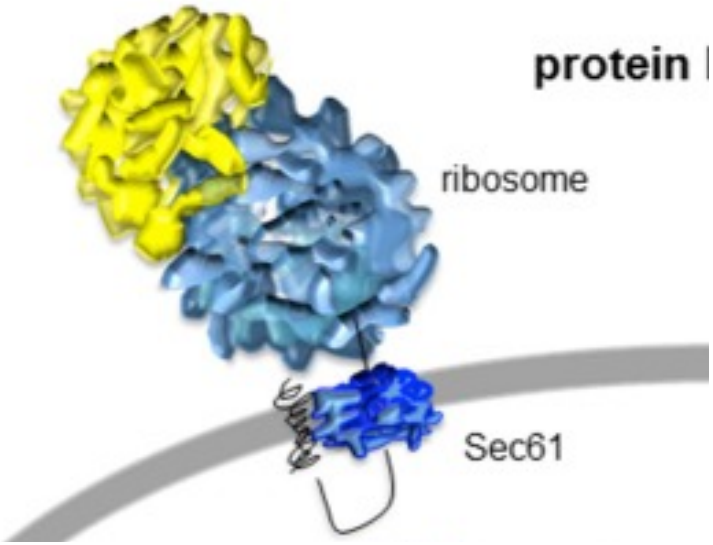
Algorithms MinDet and MinProb better under a larger MNAR ratio.

Algorithms of the first group generally seem to give better predictions.

Kim et al., Bioinformatics 21, 187 (2005)

Case study: identify clients of TRAP complex

In mammalian cells, one-third of all polypeptides are transported into or across the ER membrane via the Sec61 channel.



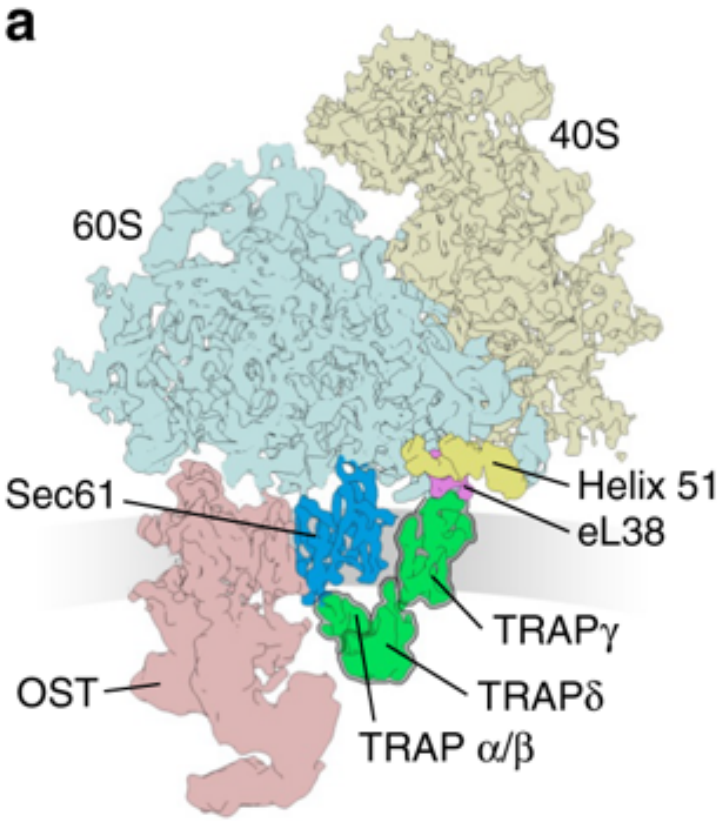
The Sec61 complex facilitates translocation of all polypeptides with signal peptides (SP) or transmembrane helices.

The Sec61-auxiliary translocon-associated protein (**TRAP**) complex supports translocation of only a **subset of precursors**.

To characterize determinants of TRAP substrate specificity, we here systematically identify TRAP-dependent precursors by analyzing cellular protein abundance changes upon siRNA-induced TRAP depletion by proteome MS.

Lang et al. Front Physiol. (2017) 8: 887

Ribosome : Sec61 : TRAP : OST supracomplex

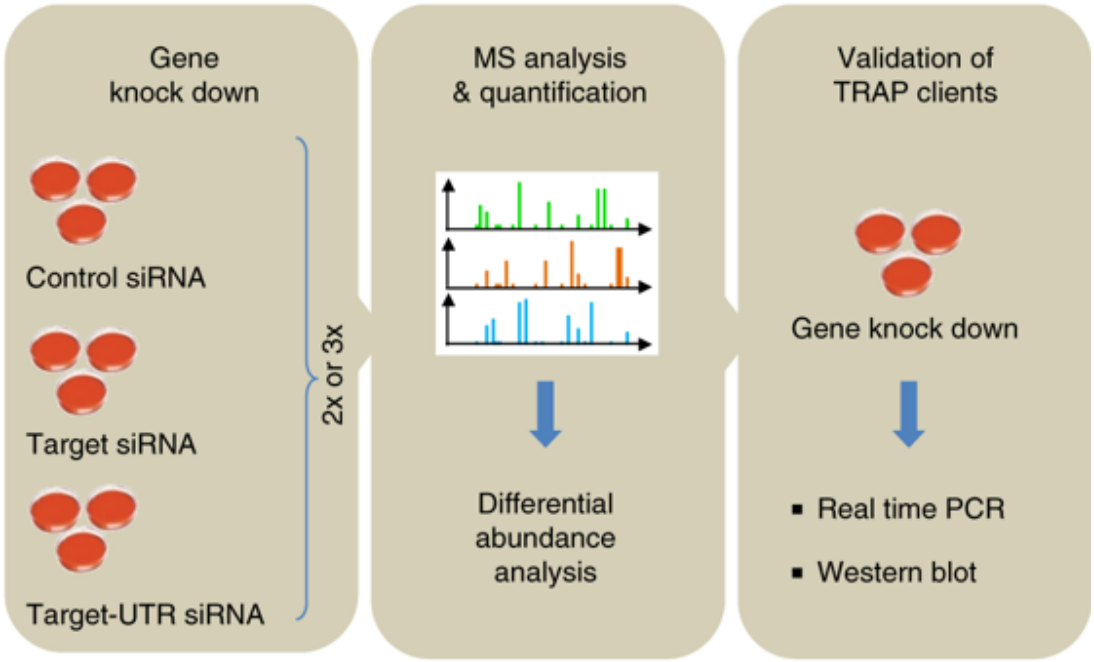


Cartoon of clipped 80S ribosome together with Sec61-complex (**blue**), TRAP-complex (**green**), and Oligo Saccharly Transferase.

Structure determined by cryo-EM

Duy et al., Nature Commun 9, 3765 (2018)

Experimental strategy



siRNA-mediated gene silencing using two different siRNAs for each target and one non-targeting (control) siRNA, respectively.

6/9 replicates for each siRNA in 2/3 independent experiments.

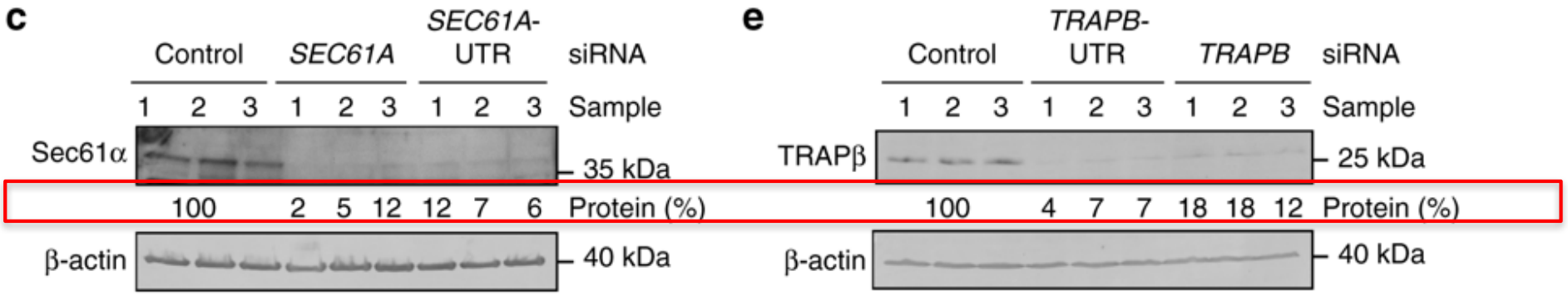
Label-free quantitative proteomic analysis and differential protein abundance analysis identify negatively affected proteins (i.e., clients) and positively affected proteins (i.e. compensatory mechanisms).

Duy et al., Nature Commun 9, 3765 (2018)

Validation of knock-down

Knock-down of SEC61-alpha

TRAP-beta



Knock-down efficiencies were evaluated by western blot.

Results are presented as % of residual protein levels (normalized to ß-actin) relative to control, which was set to 100%.

Q: why do the levels of SEC61 and TRAP do not go to zero after siRNA silencing (for 72 – 96 hours)?

Experimental strategy

Each MS experiment provided proteome-wide abundance data as LFQ intensities (Cox et al. Mol Cell Proteomics. (2014)13: 2513–2526 – how to combine peptide intensities into aggregated protein abundances?)

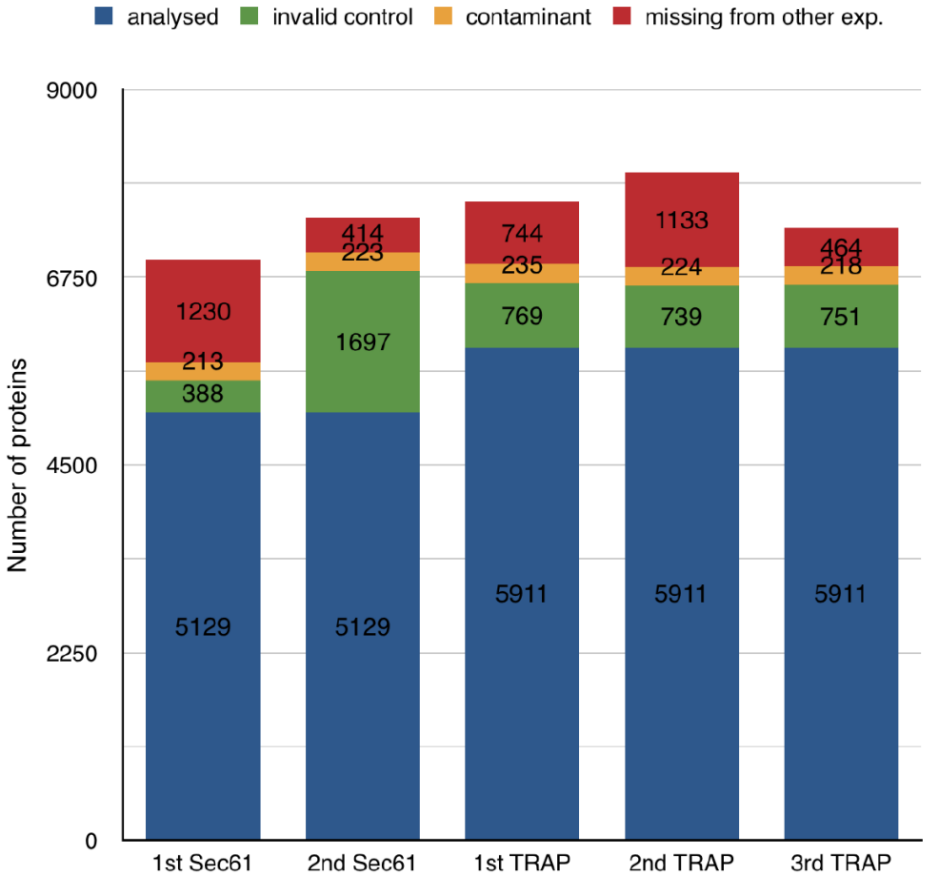
for 3 sample groups :

one control (non-targeting siRNA treated) and

two stimuli (down-regulation by two different targeting siRNAs directed against the same gene)

each having 3 data points.

Number of proteins in MS experiments



Number of proteins detected in the 2 Sec61 depletion experiments (two left most columns) and in the 3 TRAP depletion experiments (three rightmost columns).
Blue bars : proteins analyzed here.
Green : proteins that do not have sufficient control data points, i.e. more than 2/3 of the control samples have missing data points.
Yellow : “contaminants” from MaxQuant analysis.
Red : proteins that cannot be found (or contain “invalid control”) in other corresponding experiments.

The number of proteins detected in Sec61 and TRAP silencing experiments was 7212 ± 356 and 7670 ± 332, respectively (mean values with standard deviation, n=2 and n=3, respectively). The observed difference of about 460 was just a bit outside of the standard deviation and is, hence, not statistically significant.

Review comment

Reviewer #2

1. I do not understand figure 1. If only 5200/5900 proteins identified in all three experiments, why is the CV only about 300. I have the feeling that one third of the data is not reproducible. Are all regulated proteins detected in all experiments? Or do they also differ in between the experiments.

Our reply:

It is correct that a part of the data (proteins) is only detected in a fraction of the experiments.

However, our study does not claim to detect all TRAP candidates. This is not possible on the basis of the available data. Instead, we apply a conservative statistical testing scheme where only those proteins are considered as putative TRAP clients that are significantly affected by both siRNAs.

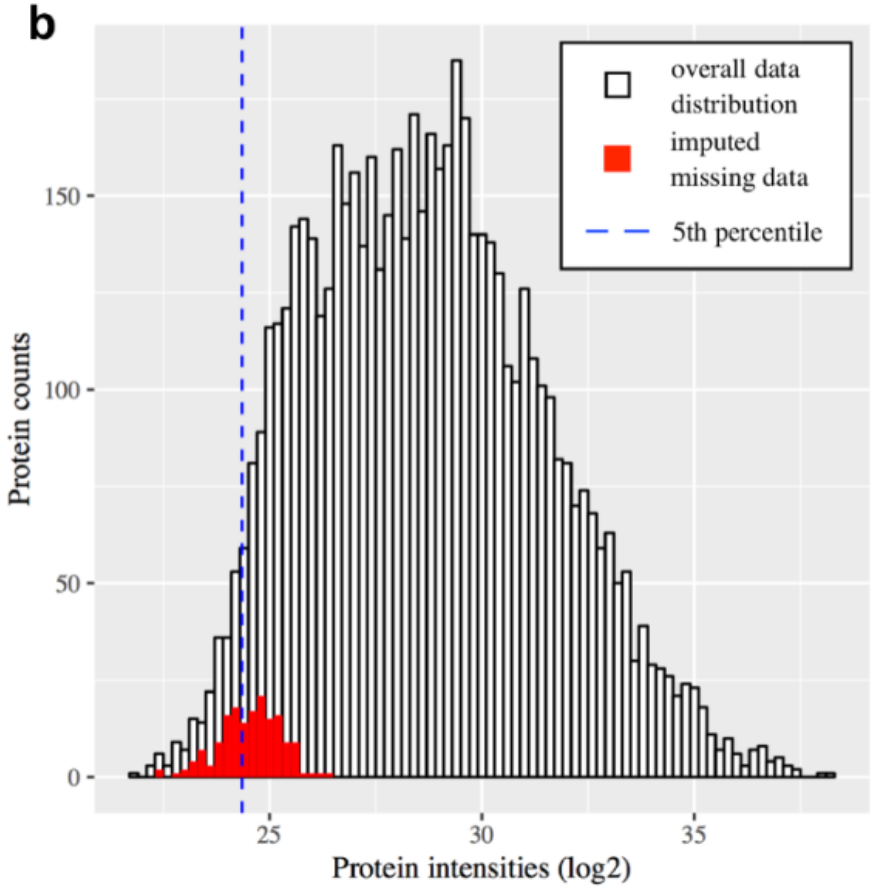
There may be further TRAP clients which either have long life times in the cell, so that they cannot be sufficiently affected by 3-4 days of siRNA silencing, or for which one siRNA was not efficient, or for other reasons.

Experimental strategy

Missing data points were generated by imputation. We distinguished 2 cases.

For **completely missing proteins** lacking any valid data points after siRNA knock-down, imputed data points were randomly generated in the bottom tail of the whole proteomics distribution.

This is based on the assumption that they come from proteins which have limited number of copies that cannot be detected by the mass spectrometry instrument.

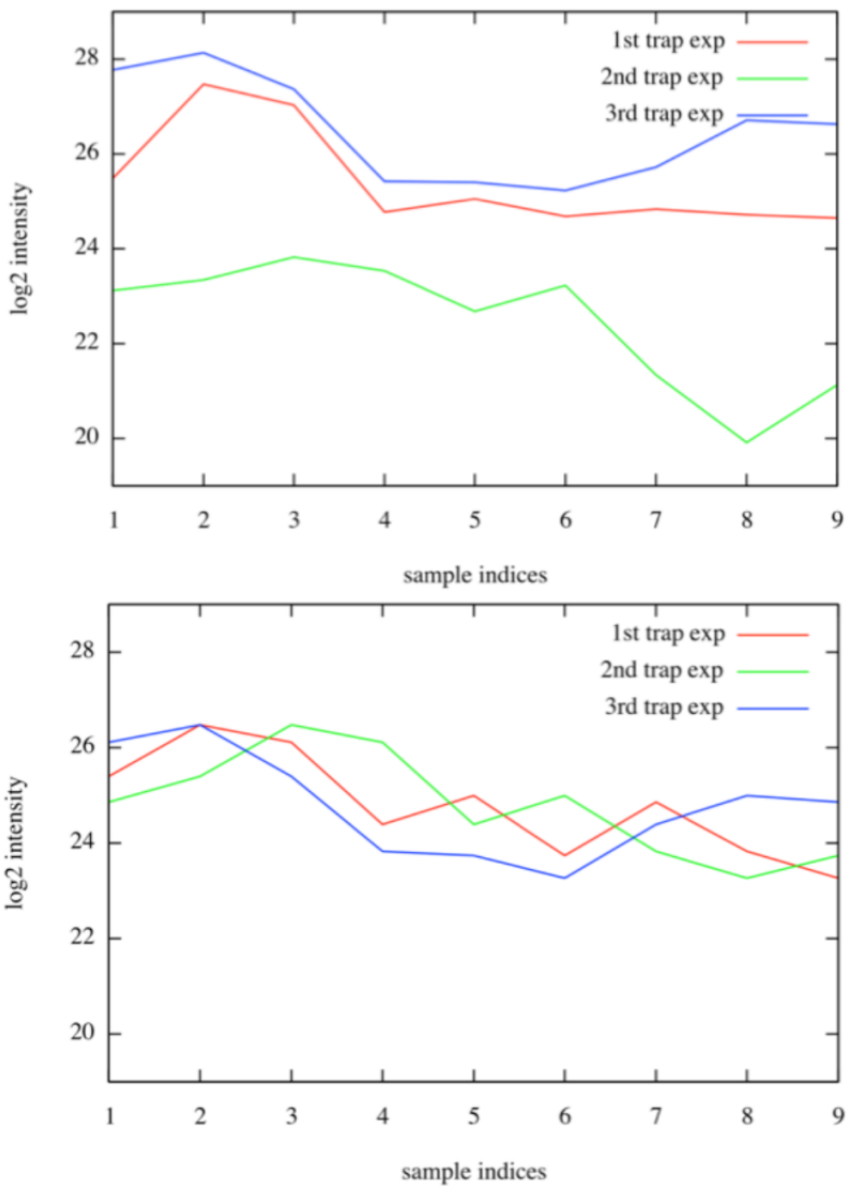


Experimental strategy

For proteins having at least one valid MS data point for knock-down samples, missing data points were generated from the valid data points based on the local least squares (LLS) imputation method (see slide 23-25 of V3).

Subsequent to data imputation, we **log2-transformed** the ratio between siRNA and control siRNA samples,
and applied **protein-based quantile normalization** to homogenize the abundance distributions of each protein with respect to statistical properties.

Protein-based quantile normalization



Intensity profiles for SSR2 protein across the 3 experiments before (top) and after (bottom) protein-based quantile normalisation.

Horizontal axis : sample IDs.

1 to 3 - control,

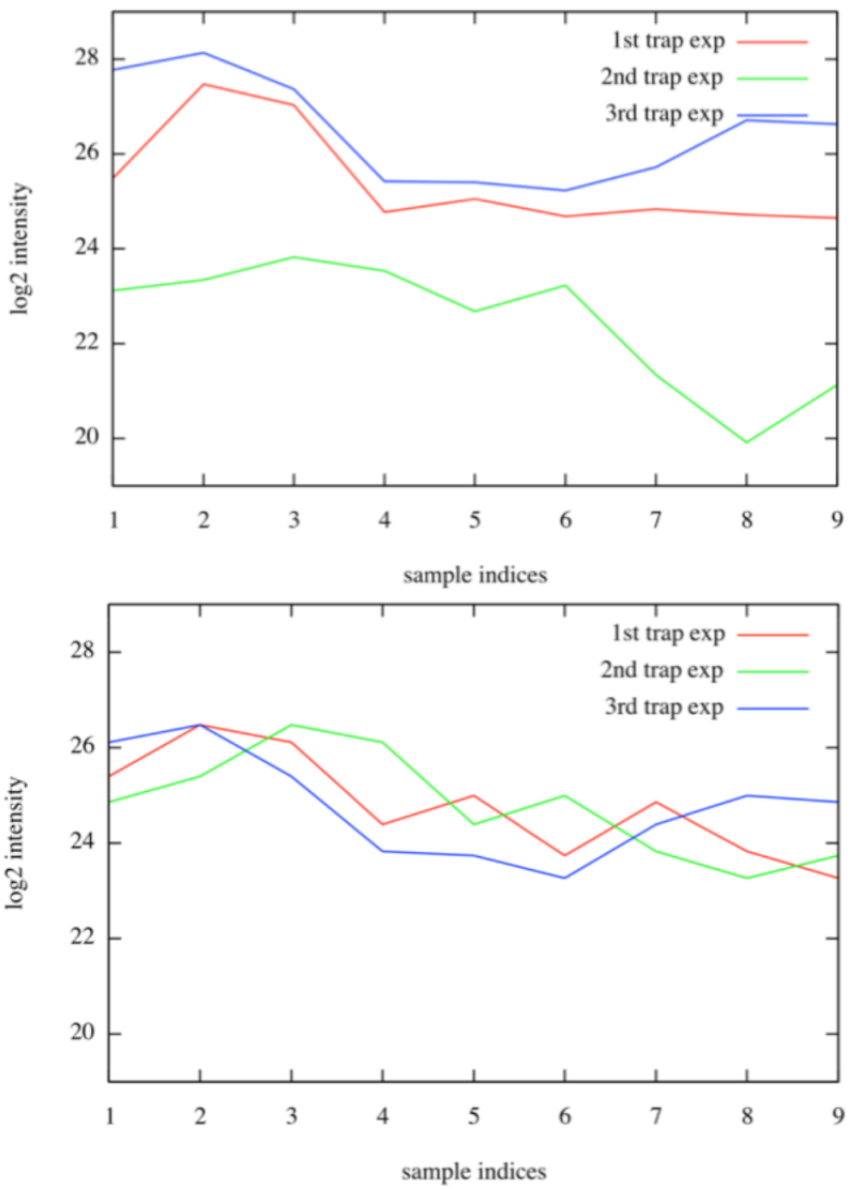
4 to 6 – SSR2 silencing by 1st siRNA,

7 to 9 – SSR2 silencing by 2nd siRNA.

Aim of protein-based quantile normalization: remove the systematic variation among 3 iterations of TRAP silencing experiment.

QN ranks the raw data, computes the averages, and replaces the original values by the ranked averages.

Protein-based quantile normalization



The variation left after normalisation reflects the biological variation between samples.

In (**a**), SSR2 levels of the controls (indices 1-3) are higher than both siRNAs in experiment 2 (red) and higher than the first siRNA in experiment 1 (blue).
In the third experiment (green), the second siRNA (indices 7-9) induces lower levels than in the controls and the first siRNA.

The same conclusions can be drawn from (**b**). The benefit of the normalized values in (**b**) is that the blue, red, and green distributions contain identical values.

Thus, one can now apply standard statistical tests to identify the significant differences.

Detection of differential abundance

Abundance in 1 siRNA knock-down was compared against control.

Proteins with an FDR-adjusted *p*-value (i.e. *q*-value) of below 5% were considered significantly affected by the siRNA knock-down.

Then, we intersected the results from the two unpaired *t*-tests for the 2 siRNAs.

This means that the abundance of all reported candidates had to be statistically significantly affected in both siRNA silencing experiments.

Review comment

6. *Statistical analysis of the data:*

On page 29 you describe imputation of data points.

Did you do a statistical analysis if the number of data points is sufficient that this imputation will not change results?

Validation of imputation method

Our reply: We assumed that … missing values … stem from “the bottom” of the distribution and belong to low abundance proteins that were not detected by the mass spectrometry instrument.

We tested to what extent the data imputation may affect the differential abundance analysis. ... The first Sec61 silencing experiment was selected for the validation… We selected only those proteins that have a “complete” dataset, i.e. none of out of nine entries was missing… This was the case for 5715 out of 6960 proteins….

To generate a synthetic dataset for missing data, we randomly removed 10% of the (known) data points from the lower tail of the distribution ...

For two different thresholds (5th and 10th percentile of the overall distribution), we repeated the removal 100 times. Therefore, in total, we generated 200 new datasets with artificially generated “missing” data.

Validation of imputation method

Subsequently, these “missing” data points were imputed.

Then, a differential protein abundance analysis was carried out on the imputed and the original data.

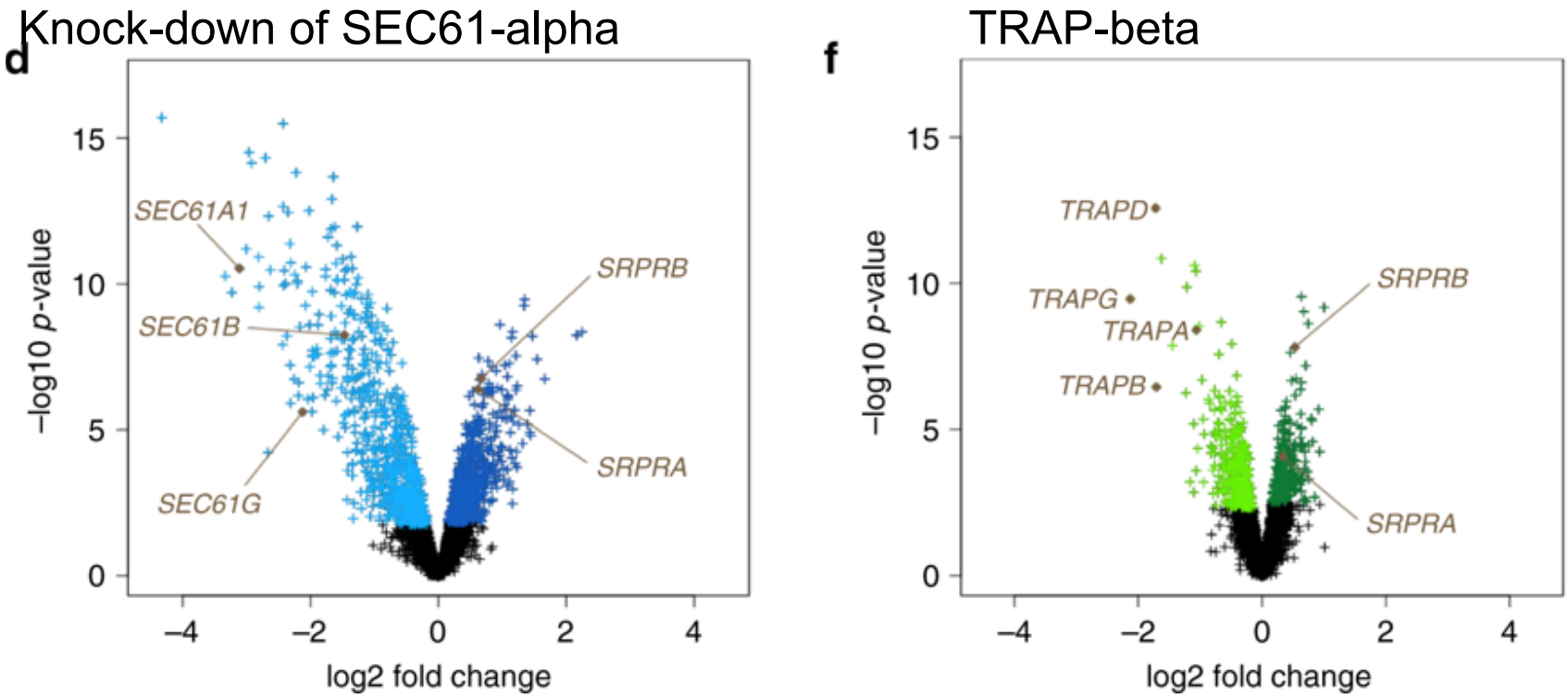
Finally, we compared the results of the differential analysis of the imputed and original data to validate the reliability of the imputation method.

For this, using the results of the previous steps, the significantly affected proteins were either labelled as 1 (positively affected) or as -1 (negatively affected) while the unaffected proteins were labelled 0.

Afterwards, we computed the Pearson correlation coefficient between the results of the original data and of the imputed data.

The overall correlation coefficients for the 5th and 10th percentile thresholds are 0.975 ± 0.018 and 0.927 ± 0.020, respectively.

Volcano plot: differential protein abundance



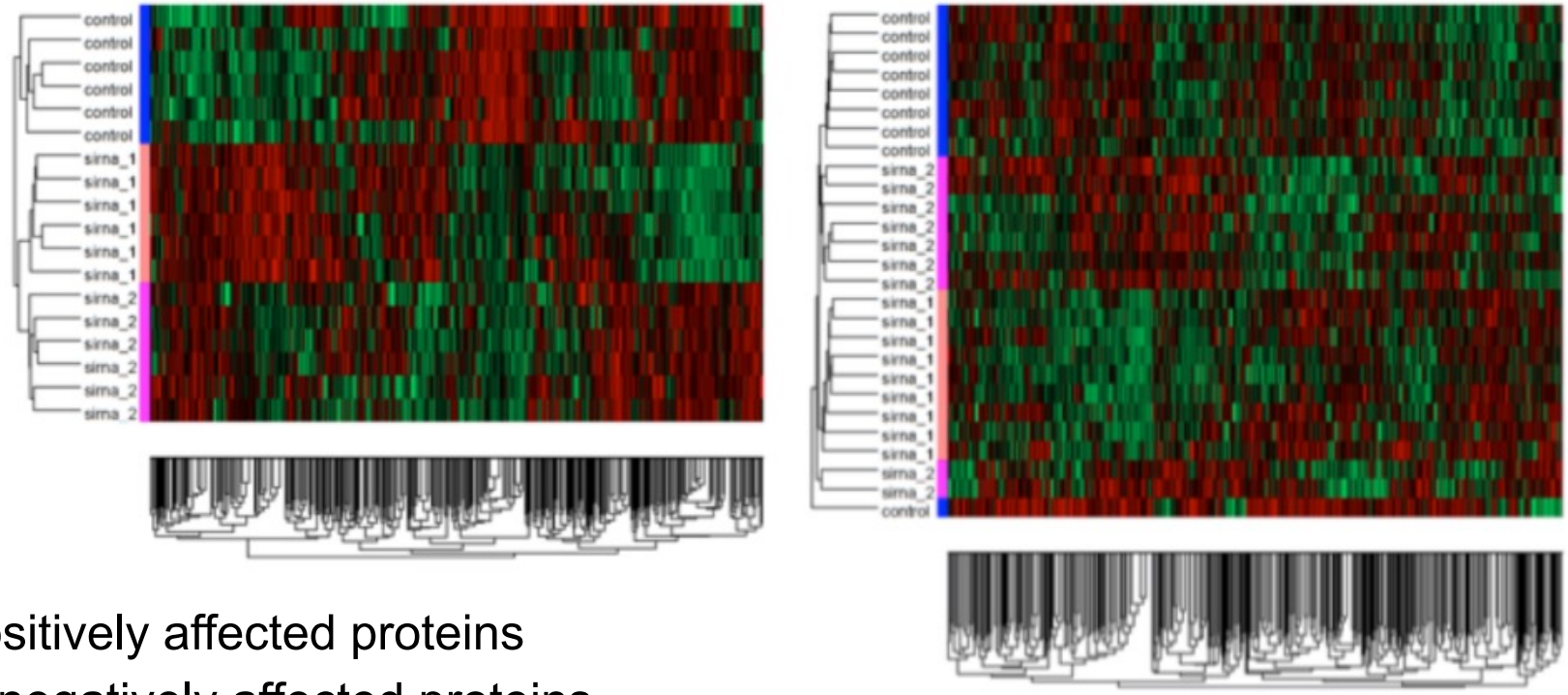
Differentially affected proteins were characterized by the mean difference of their intensities plotted against the respective permutation false discovery rate-adjusted *p*-values in volcano plots.

The results for a single siRNA are shown in each case (*SEC61A1*-UTR siRNA, *TRAPB* siRNA).

Up- / down-regulation

Heat maps visualize clusters of proteins that were

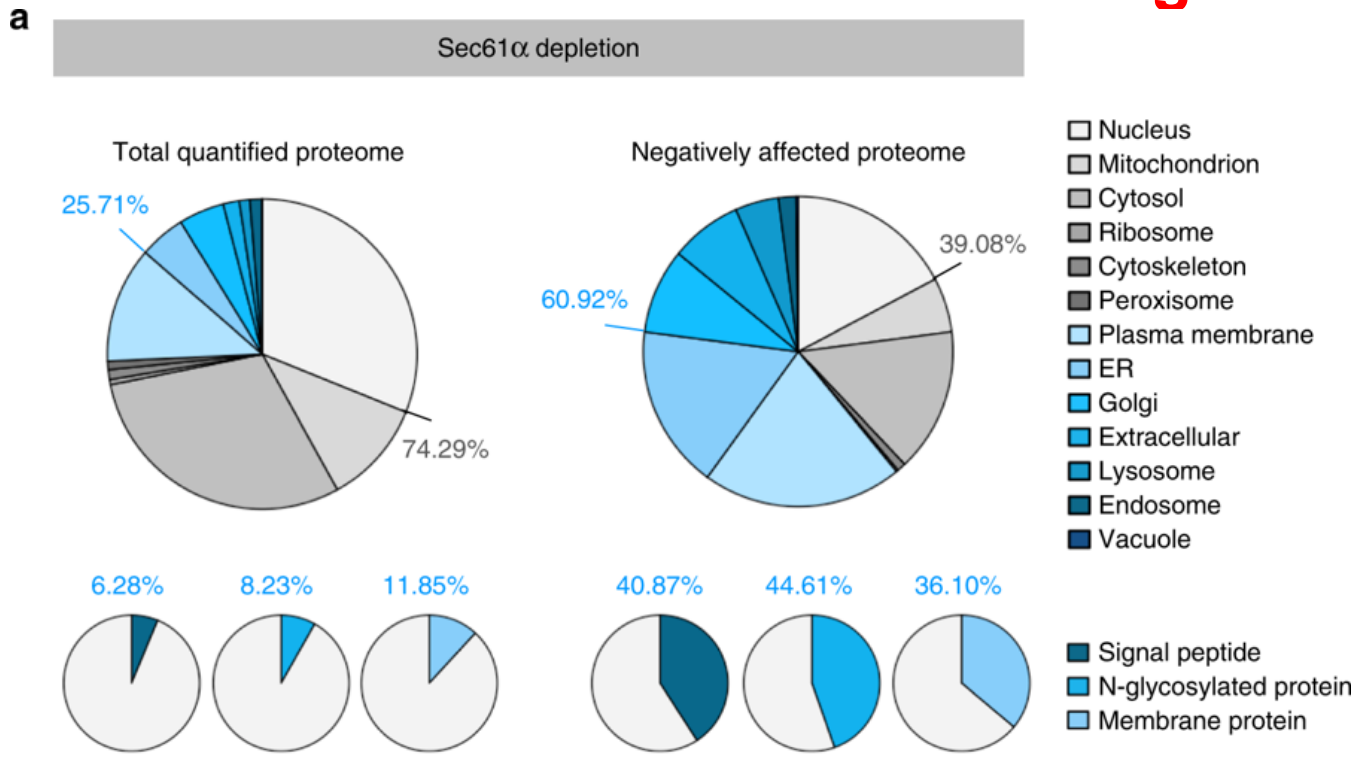
- significantly **upregulated** following treatment with both siRNAs directed against either *SEC61A1* (left) or *TRAPB* (right) mRNA or with non-targeting (control) siRNA, or that were
- significantly **downregulated**, or that
- represent **variations** between siRNAs.



Red : positively affected proteins

Green : negatively affected proteins.

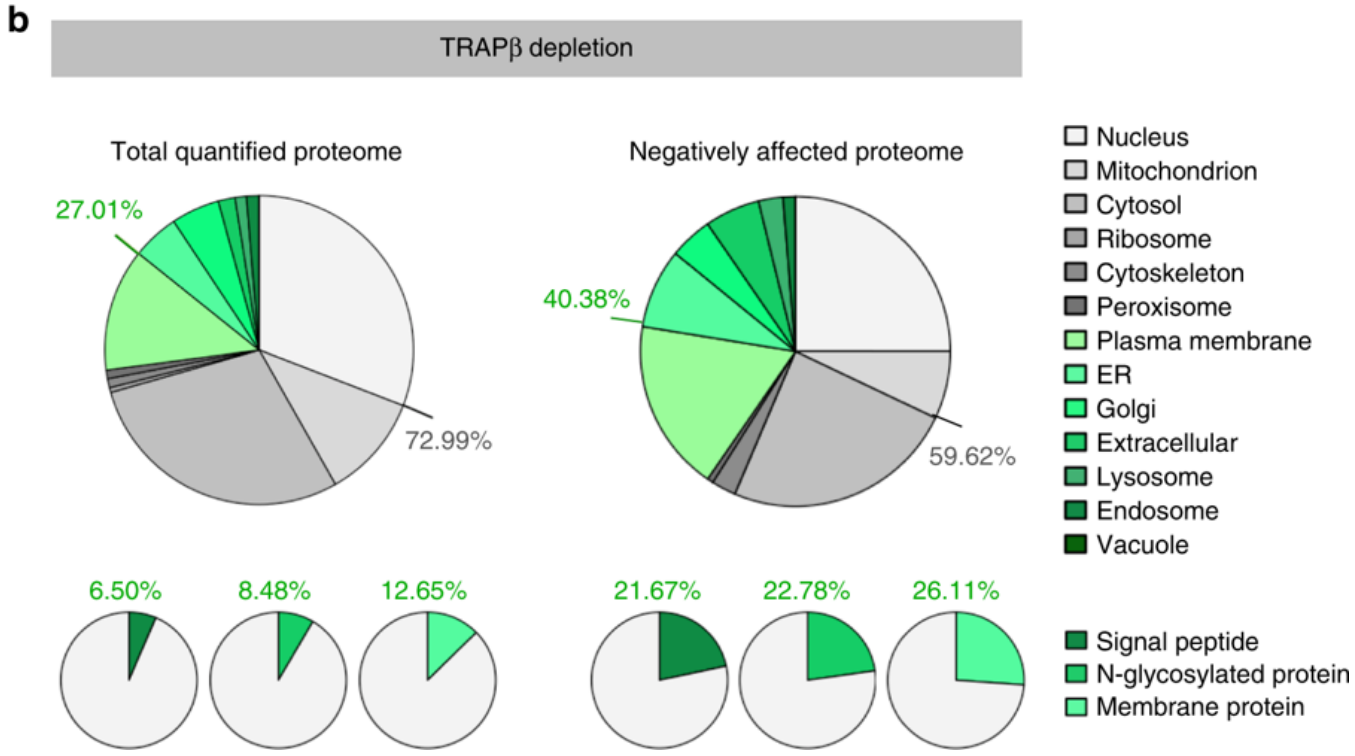
Annotation of differentially abundant proteins after Sec61 silencing



Validation of Sec61 clients based on Gene Ontology enrichment factors.

Protein annotations of signal peptides, membrane location, and N-glycosylation in humans were extracted from UniProtKB, and used to determine the enrichment of Gene Ontology annotations among the secondarily affected proteins.
Summary of two Sec61 depletion experiments performed in triplicate.

Annotation of differentially abundant proteins after TRAP silencing

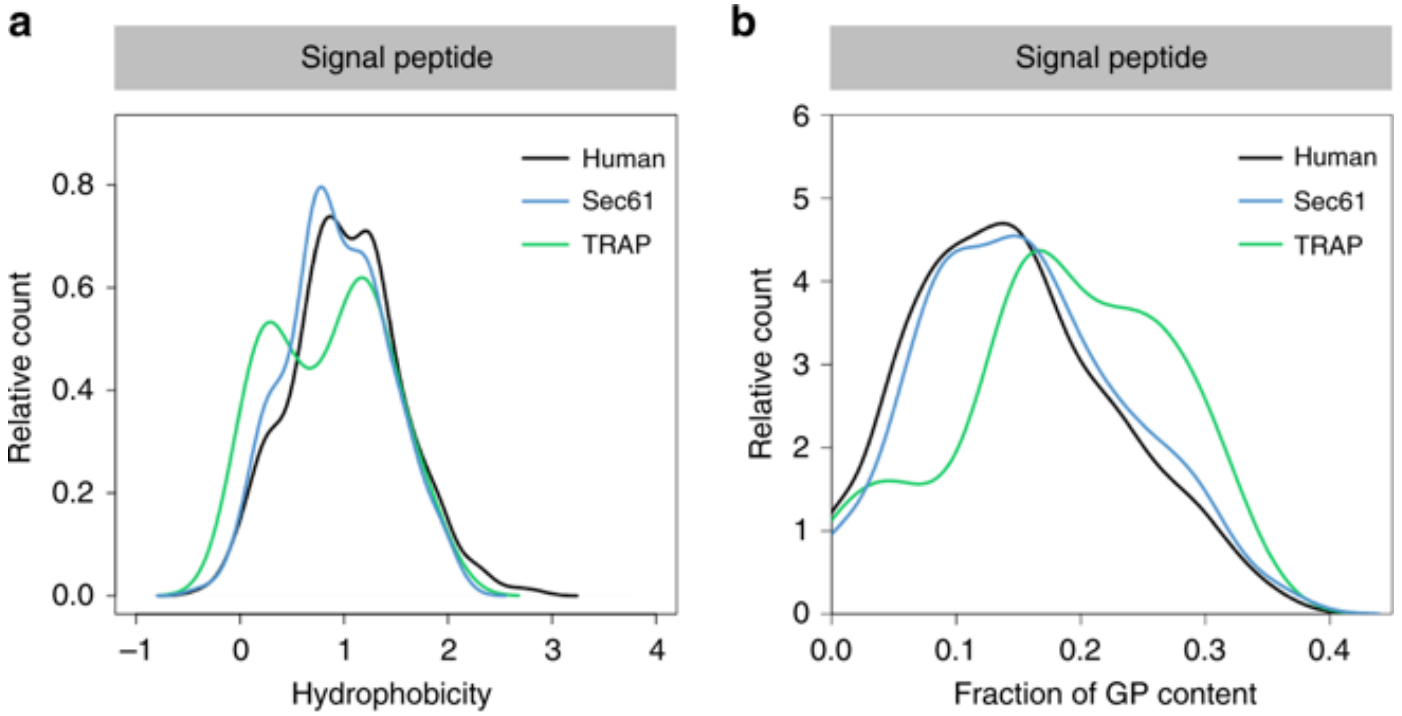


Validation of TRAP clients based on Gene Ontology enrichment factors.

Summary of three TRAP depletion experiments performed in triplicate.

→ clear enrichment of green fraction (ER targeted organelles)

Physicochemical properties of TRAP clients



The signal peptides of TRAP clients are **less hydrophobic** and have a **higher Gly/Pro content** than Sec61 clients and the full proteome.

Physicochemical properties of TRAP clients with SP.

Hydrophobicity score (**a**) and glycine/proline (GP) content (**b**) of SP sequences. Hydrophobicity score was calculated as the averaged hydrophobicity of its amino acids according to the Kyte-Doolittle propensity scale. GP content was calculated as the total fraction of glycine and proline in the respective sequence.