Supplementary Text: Mathematical formulation of the Subcellular Spatial Razor.

A. The subcellular spatial razor.

The subcellular spatial razor model assumes that a given protein can be in the nucleus (n) and in the cytoplasm (c) for both unstimulated (u) cells and stimulated (s) cells. The corresponding abundances (Fig S1) are $A_{n,u}$, $A_{n,s}$, $A_{c,u}$, $A_{c,s}$. There are three SILAC ratios given by:

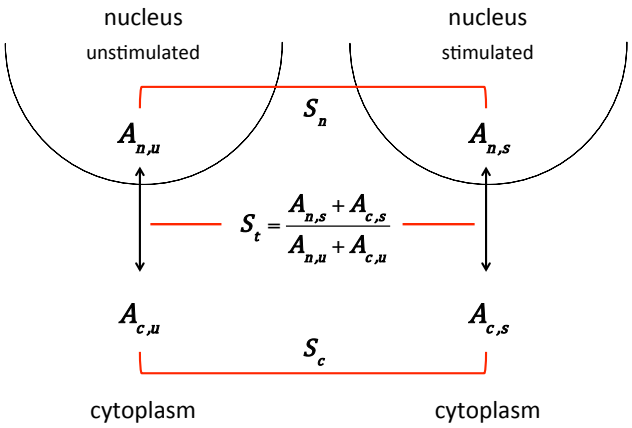


Fig. S1. The spatial razor model.

$$S_n = A_{n,s}/A_{n,u}, \qquad [1]$$
$$S_c = A_{c,s}/A_{c,u},$$
$$S_t = (A_{n,s} + A_{c,s})/(A_{n,u} + A_{c,u}).$$

The values for the fractions of the protein in the nucleus in unstimulated and stimulated cells, f_u and f_s, are closely related to the SILAC ratios:

$$f_u = A_{n,u}/(A_{n,u} + A_{c,u}) = (S_t - S_c)/(S_n - S_c), \qquad [2]$$
$$f_s = A_{n,s}/(A_{n,s} + A_{c,s}) = S_n(S_t - S_c)/S_t(S_n - S_c).$$

Furthermore, the fractions f_u and f_s are closely related to the parameters used in a three-dimensional orthogonal basis set for the experimental results (see below).

$$S_n/S_t = f_s/f_u, \qquad [3]$$
$$S_c/S_t = (1 - f_s)/(1 - f_u)$$

Experimental application of the spatial razor requires fractionating the total cellular proteins (total fraction) into two subcellular fractions such that: nucleus-enriched fraction + nucleus-depleted fraction (cytoplasm) = total fraction and measuring the set of SILAC ratios $\{S_n, S_c, S_t\}$. Although formulated here for the nucleus, a spatial razor could also be applied to other subcellular locations through the use of sample triplets such as {mitochondria-enriched, mitochondria-depleted, total-fraction}.

B. An orthogonal basis set.

Because changes in both the total abundance and the subcellular distribution of a protein influence its abundance in the nucleus/cytoplasm, the set of SILAC ratios $\{S_n, S_c, S_t\}$ are not an orthogonal basis set. As we have shown previously,[1] the 3D orthogonal basis set $\{S_n/S_t , S_c/S_t , S_t\}$ separates changes in total protein abundance (S_t) from changes in nucleus/cytoplasm distribution (the $\{S_n/S_t , S_c/S_t\}$ distribution plane). In this distribution plane, for any fraction of the protein f_u in the nucleus in the unstimulated cells, a unique curve is obtained as the fraction of the protein in the nucleus in the stimulated cells is varied over $0 < f_s < 1$,

independent of changes in total abundance (Fig. S2). The origin of the plot corresponds to f_u/f_s = 1, i.e. to no change in distribution upon stimulation of the cells. Conservation of mass requires that the data points lie in the two indicated quadrants of the distribution plane, which correspond to N ➔ C and C ➔ N redistribution of the subcellular location following cellular stimulation.

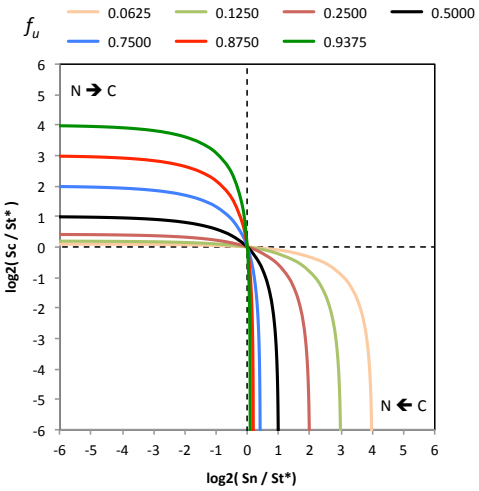


Fig. S2. For different values of f_u, the location in the distribution plane as f_s is varied over $0 < f_s < 1$.

C. Correction for Enrichment during MS Sampling.

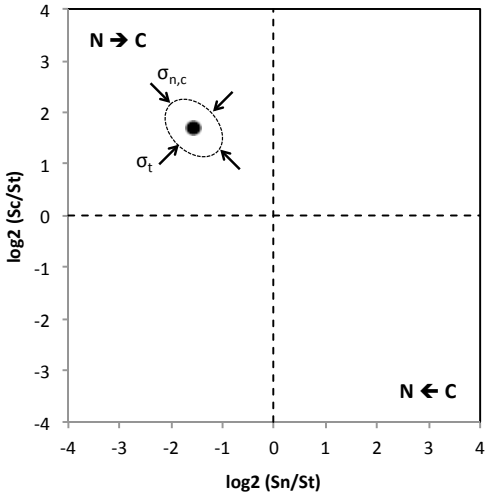
As we have shown elsewhere,[2] joint co-processing of the C and N MS data sets (a C&N data set) allows calculation of an apparent change in total abundance S_t*. A subtlety is that usually equal amounts of N and C proteins are taken for the MS data collection. This amounts to enrichment of the nuclear proteins relative to the cytoplasmic proteins during the data collection, e.g. in the present experiments 30 µg of nuclear proteins and 30 µg of cytoplasmic proteins were used in the MS runs. A global nuclear enrichment factor during data collection can be calculated from the overall yields of protein in the nuclear and cytoplasmic fractions. Using this known enrichment factor (r), we have shown elsewhere[2] that the measured value of R_t can be used to obtain a corrected change in abundance S_t^* with the equation:

$$S_t^* = R_t \left\{ \frac{\left(\frac{S_n}{R_t}-1\right)+\left(1-\frac{S_c}{R_t}\right)}{r\left(\frac{S_n}{R_t}-1\right)+\left(1-\frac{S_c}{R_t}\right)} \right\} \left\{ \frac{r\frac{S_c}{S_n}\left(\frac{S_n}{R_t}-1\right)+\left(1-\frac{S_c}{R_t}\right)}{\frac{S_c}{S_n}\left(\frac{S_n}{R_t}-1\right)+\left(1-\frac{S_c}{R_t}\right)} \right\}, \qquad [4]$$

where R_t is the apparent total SILAC ratio obtained from joint processing of the C and N data sets. As shown elsewhere,[2] the correction factor is symmetric with respect to enrichment/depletion of one fraction relative to the other. For $f_s = f_u$ (no redistribution), $S_t^* = R_t$, i.e. there is no distortion of S_t^* for any value of f_u. This is because in the absence of subcellular redistribution, the conservation of mass equation $S_t = S_n = S_c$ applies. The magnitudes of the corrections can become appreciable only for large/small global enrichment factors r coupled with substantial redistribution ($0.5 \gtrsim f_s/f_u \gtrsim 2$), even if there are large changes in total abundance with $1 \ll S_t$ or $1 \gg S_t$.[2] In the present experiments, the correction factor ranged over 0.3 – 1.9 with the large majority of proteins close to 1. This correction in the estimate of S_t* does not affect S_n/S_c and only affects the evaluation of whether the protein abundance is predominantly skewed to the nucleus or cytoplasm (see Fig. S3 below).

D. Evaluation of Significance.

A more complete development of the statistics for identification of outliers with significance in the spatial razor framework will be given elsewhere. Here we note that for those proteins detected in both the N and C compartments, because of the coupling between S_n and S_c there are two degrees of freedom in the spatial razor, which can be expressed in terms of the nuclear fractions f_s and f_u or of the measured quantities S_t and S_n/S_c. When the data is presented in the 3D space $\{S_n/S_t\,, S_c/S_t\,, S_t\}$, experimental variation in S_t is expressed both along the S_t axis as well as in the distribution plane parallel to the line S_c/S_t = S_n/S_t (Fig. S3). There is no *a priori* reason to assume that changes in abundance and changes in subcellular location are correlated. Therefore for each protein "significance" for changes in S_t and S_n/S_c should be evaluated independently. To avoid mixing independent parameters, in the main text the positioning of the data points in the 3D space uses S_t* in the distribution plane and S_t for the color-coded abundance axis.



. S3. The standard deviations of S_n/S_c and S_t in the 3D space Fig$\{S_n/S_t\,, S_c/S_t\,, S_t\}$.

E. Screening the Data for MS Sampling and Subcellular Fractionation/Extraction Artefacts.

A fundamental assumption in all quantitative proteomics experiments that compare different cellular states is that, for a given sample type, the efficiency of extraction for the individual proteins does not vary with cellular state. This assumption is explicit in SILAC and related differential labelling approaches, but is also implicit in other types of experimental strategies such as the monitoring with MS intensities of changes in abundance of unlabelled single reaction products.

On the other hand, it is well known that protein-polynucleotide interactions complicate complete extraction of proteins from nuclei[3] and the same is probably to be expected for mitochondria. If a protein traffics between the cytoplasm and the nucleus, and, for example, because of interactions with chromatin becomes more difficult to extract from the nucleus, this may then vary the extractability of the protein between stimulated/unstimulated cells. For total lysate (T) samples, this is not *a priori* distinguishable from total abundance changes and trafficking effects may be misinterpreted as abundance changes.

The spatial razor offers a framework for using conservation of mass to detect potential problems of this kind. The explicit inclusion of conservation of mass implies that for proteins present in a single compartment, S_n/S_t = 1 (nucleus) or

S_c/S_t =1 (cytoplasm). A similar principle applies to proteins present in both compartments: for ideal MS sampling and extraction/fractionation, joint co-processing of the C and N MS data sets should give an apparent change in abundance S_t^* such that S_t^*/S_t = 1. All of the ratios S/S_t gave essentially Gaussian distributions, as shown in Fig. S4 for S_t^*/S_t. The outliers from S_t^*/S_t = 1 tend to be proteins with fewer ratio counts that were quantified in fewer replicates. To select a high-confidence set of proteins we therefore restricted the 134-set of proteins to those with $|\log_2(S/S_t)|$ < 0.5 that were quantified with at least 10 SILAC ratio counts in at least 2 replicates.

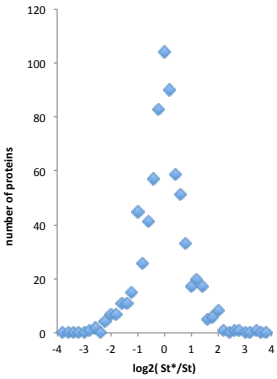
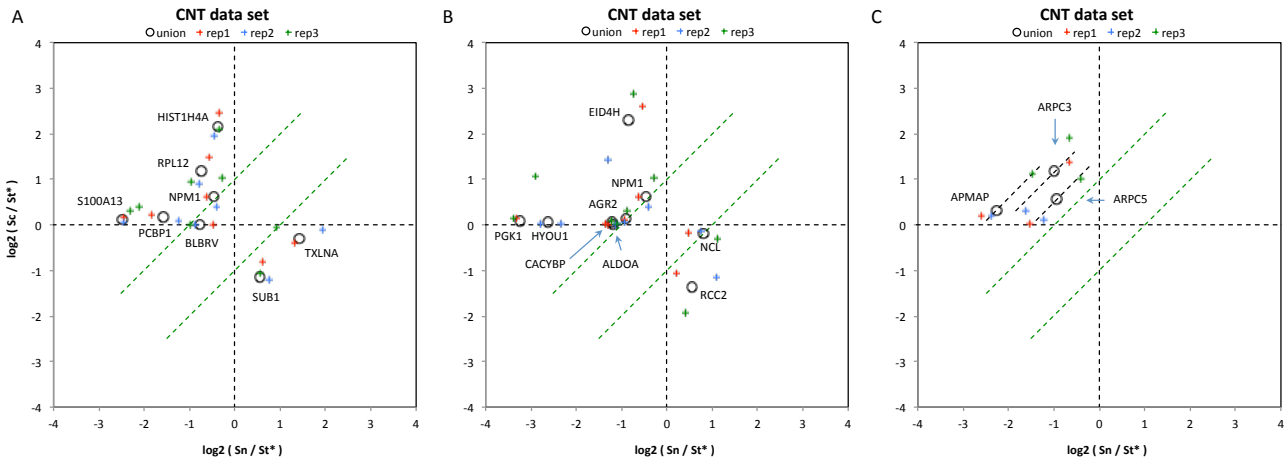


Fig. S4. The scatter of S_t^*/S_t used in the selection of high confidence proteins.

With continuing improvements in MS instrumentation, reproducible, reliable sample preparation may become a limiting factor for quantitative proteomics and we therefore looked in more detail at the scatter over replicates in the present experiments. For the 134-set of proteins, typical scatter over the three replicates is shown in Fig. S4. The union is the average weighted by the MS intensity recorded in each replicate. The scatter varies with the protein, with a tendency for greater scatter for those proteins that show large changes in S_c/S_t and/or S_n/S_t (Fig. S5 A, B). In general the behaviour of the individual proteins is clear despite the scatter over the replicates, which indicates adequate consistency with conservation of mass.



Fig, S5. (A-C) 3D subcellular spatial razor plots for various proteins of the 134-set.

The nature of the scatter in Fig. S5 indicates that it may be caused by both variation in cellular response and/or subcellular fractionation ($\sigma_{n,c}$ in Fig. S3) and by recovery of proteins during extraction (σ_t in Fig. S3). In fact, MS sampling and/or extraction/recovery tended to be a limiting factor for some proteins, which showed small scatter in S_n/S_c and larger scatter for S_t^* (Fig. S5 C). We suspect that many of the proteins that translocate interact with chromatin and that all samples should be exhaustively extracted serially using procedures appropriate for recovery of proteins from protein/polynucleotide complexes (high salt, DNAse and possibly RNAse), including the cytoplasmic sample because of the presence of DNA in

mitochondria. Because complete, reproducible protein extraction is also crucial for proteomics measurements of total abundances (S_t) in many other contexts, we are currently investigating whether this improves the verification of conservation of mass for less abundant proteins and therefore allows greater proteome coverage. If adequate consistency with conservation of mass is demonstrated for the MS sampling/fractionation/extraction protocols, lower abundance proteins can be interpreted with greater confidence and in principle the fraction of a protein in the nucleus/cytoplasm for unstimulated/stimulated cells can be evaluated with the spatial razor (Eqns. [2]).

Supplement References.

1. Mulvey CM, Tudzarova S, Crawford M, Williams GH, Stoeber K & Godovac-Zimmermann J. (2013) Subcellular proteomics reveals a role for nucleo-cytoplasmic trafficking at the DNA replication origin activation checkpoint. J Proteome Res, 12, 1436-53.

2. Baqader N, Radulovic M, Mulvey CM, Crawford M, Williams GM, Stoeber K, and Godovac-Zimmermann J. Comparison of Subcellular Nuclear-Cytoplasmic Abundance and Trafficking of Proteins for Human Fibroblasts under Oxidative Stress or G1/S Cell Cycle Arrest. *in preparation*.

3. Torrente MP, Zee BM, Young NL, Baliban RC, LeRoy G, et al. (2011) Proteomic Interrogation of Human Chromatin. PLoS ONE 6(9): e24747. doi:10.1371/journal.pone.0024747