nature COMMUNICATIONS

ARTICLE

Received 11 Feb 2015 | Accepted 22 Oct 2015 | Published 12 Jan 2016

DOI: 10.1038/ncomms9992

OPEN

A draft map of the mouse pluripotent stem cell spatial proteome

Andy Christoforou[1,2], Claire M. Mulvey[1,2], Lisa M. Breckels[1,3], Aikaterini Geladaki[1,2], Tracey Hurrell[1,4], Penelope C. Hayward[2], Thomas Naake[1,3], Laurent Gatto[1,3], Rosa Viner[5], Alfonso Martinez Arias[2] & Kathryn S. Lilley[1]

Knowledge of the subcellular distribution of proteins is vital for understanding cellular mechanisms. Capturing the subcellular proteome in a single experiment has proven challenging, with studies focusing on specific compartments or assigning proteins to subcellular niches with low resolution and/or accuracy. Here we introduce hyperLOPIT, a method that couples extensive fractionation, quantitative high-resolution accurate mass spectrometry with multivariate data analysis. We apply hyperLOPIT to a pluripotent stem cell population whose subcellular proteome has not been extensively studied. We provide localization data on over 5,000 proteins with unprecedented spatial resolution to reveal the organization of organelles, sub-organellar compartments, protein complexes, functional networks and steady-state dynamics of proteins and unexpected subcellular locations. The method paves the way for characterizing the impact of post-transcriptional and post-translational modification on protein location and studies involving proteome-level locational changes on cellular perturbation. An interactive open-source resource is presented that enables exploration of these data.

[1] Department of Biochemistry, Cambridge Centre for Proteomics, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK. [2] Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK. [3] Department of Biochemistry, Computational Proteomics Unit, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK. [4] Department of Pharmacology, University of Pretoria, Arcadia 0007, Republic of South Africa. [5] Thermo Fisher Scientific, 355 River Oaks Pkwy, San Jose, California 95314, USA. Correspondence and requests for materials should be addressed to K.S.L. (email: k.s.lilley@bioc.cam.ac.uk).

Pluripotent mouse embryonic stem (ES) cells are self-renewing clonal populations derived from blastocysts, which can be differentiated into the ensemble of cell types of the organism[1,2]. Their study is central to developmental biology and the emerging field of regenerative medicine. Currently our understanding of the biology of ES cells is deeply rooted in our knowledge of their transcriptomes, epigenetics and underlying gene regulatory networks, which have created a foundation for understanding pluripotency and the transition to differentiation[3,4].

There is evidence that post-transcriptional events such as signalling, adhesion, protein turnover and post-translational modification make a significant contribution to the regulation of differentiation[5–8], yet their precise roles in this process, and how they interact with each other and with the transcriptional machinery, remain open questions. The transition from self-renewal to differentiation is also associated with major changes in cell morphology, and therefore some of the effects of these post-transcriptional processes must be associated with changes in intracellular organization. Understanding the subcellular distribution of proteins and other biomolecules, and how the distribution changes with cell state, is therefore of paramount importance for the delineation of post-transcriptional processes in ES cells.

Protein localization is typically determined by immunocytochemistry or by monitoring fluorescent fusion proteins by confocal microscopy. While these approaches are valuable and well-established, there are certain limitations to their applicability. Immunocytochemistry is dependent on the availability of high-specificity and high-sensitivity antibodies, while fluorescent fusion proteins are vulnerable to aberrant localization due to the effect of the fusion moiety on protein topology[9,10]. These limitations can be overcome with complementary technologies such as protein mass spectrometry (MS), which offers the capability to assay thousands of proteins simultaneously and in their native state[11].

Localization of organelle proteins by isotope tagging (LOPIT) is a quantitative proteomics method for the high throughput and simultaneous characterization of multiple subcellular compartments, without the requirement for total purification of compartments of interest[12]. LOPIT combines biochemical fractionation by density-gradient ultracentrifugation, sample multiplexing by *in vitro* covalent labelling, and liquid chromatography-mass spectrometry. In LOPIT, cells are first lysed under detergent-free conditions so that there is minimal disruption to organelle integrity. Membranes are then separated based on their characteristic buoyant densities by ultracentrifugation. Although organelles do not partition into discrete purified fractions, different organelles display distinct enrichment patterns. Fractions representing peak enrichment for organelles of interest are selected for proteolytic digestion. The resulting peptides are differentially labelled with amine-reactive tandem mass tag (TMT) reagents[13], which allow peptides derived from each fraction to be distinguished by mass spectrometry. The relative abundance of a peptide can be determined by its TMT reporter ion profile, which recapitulates the distribution of the protein across the fractionation scheme. Proteins residing in the same subcellular niche would be expected to co-distribute, and therefore present similar TMT reporter ion profiles[14]. Classification algorithms are then used to assign proteins to subcellular compartments based on correlation with organellar marker proteins.

Here we significantly extend the LOPIT concept with novel approaches for sample preparation, mass spectrometry data acquisition and multivariate analysis. This new workflow, named hyperplexed LOPIT (hyperLOPIT), benefits from several recent technological advancements. First, the development of neutron-encoded isotopologue variants of TMT has increased the multiplexing capacity of isobaric tagging experiments to 10 samples[15]. These additional labels have enabled more subcellular fractions to be sampled, allowing for a more elaborate fractionation scheme that reaches sub-organellar levels of resolution. Second, quantitative accuracy of TMT-based applications is significantly improved by mass spectrometry data acquisition using synchronous precursor selection MS^3 (SPS-MS^3; see Supplementary Fig. 1 for a detailed overview of this method). Multivariate approaches such as hyperLOPIT represent particularly demanding applications of TMT quantification, as consistently high accuracy and precision are necessary for co-localized proteins to display correlated TMT distributions[16]. We have therefore incorporated SPS-MS^3 acquisition on the Orbitrap Fusion Tribrid mass spectrometer (Thermo Fisher Scientific) into the hyperLOPIT pipeline, and demonstrate that it greatly improves spatial resolution and the reliability with which protein localization may be determined. Finally, we have extended our data analysis platform to facilitate rapid interrogation of the data by providing an easy-to-use graphical user interface.

We apply the hyperLOPIT workflow to a population of self-renewing mouse ES cells. The result is comprehensive coverage of the subcellular proteome with unprecedented spatial resolution, enabling the mapping of components of organelles, transitory proteins, multi-protein complexes, signalling pathways and families of functionally related proteins. The presented data offer hitherto unknown subcellular detail about a population of self-renewing stem cells. We observe the dynamic nature of Golgi apparatus proteins, noting sets of Golgi marker proteins that are distributed amongst other subcellular structures, an observation we support by microscopy. We also demonstrate that additional insights into published interactomes and focused protein-localization studies can be attained by integrative analysis with LOPIT. HyperLOPIT offers a spatial scaffold onto which other high content proteomics data sets can be mapped yielding added value to complementary data. We give an example of this functionality by combining the data presented here with a recent data set of surface proteins captured using chemical tagging.

Results

Biochemical fractionation of mouse ES cells. An overview of the hyperLOPIT workflow is shown in Fig. 1. To create a method fit for the purpose of capturing cell-wide proteome localization data, we first improved the LOPIT workflow as follows: to increase subcellular resolution, extensive subcellular fractionation of pluripotent E14TG2a cells was performed. First, crude membranes were separated from the soluble fraction, enriched in cytosolic proteins. Crude membranes were then separated into organelle-enriched fractions by equilibrium density gradient centrifugation, within which organelles adopt specific distribution profiles consistent with their respective buoyant densities. A portion of the cell culture was also used to enrich chromatin-associated proteins with a parallel workflow based on detergent permeabilization[16]. Ten fractions were chosen to best represent peak organelle densities, labelled with TMT reagents and processed as described in the Methods section. We acquired three biological replicates of hyperLOPIT data, each with a slightly different selection of subcellular fractions for TMT 10-plex labelling. The first replicate placed greater emphasis on resolving low density-gradient fractions that are enriched in secretory pathway components, while the second and third replicates placed greater emphasis on separation of the denser organelles such as the mitochondrion and peroxisome.

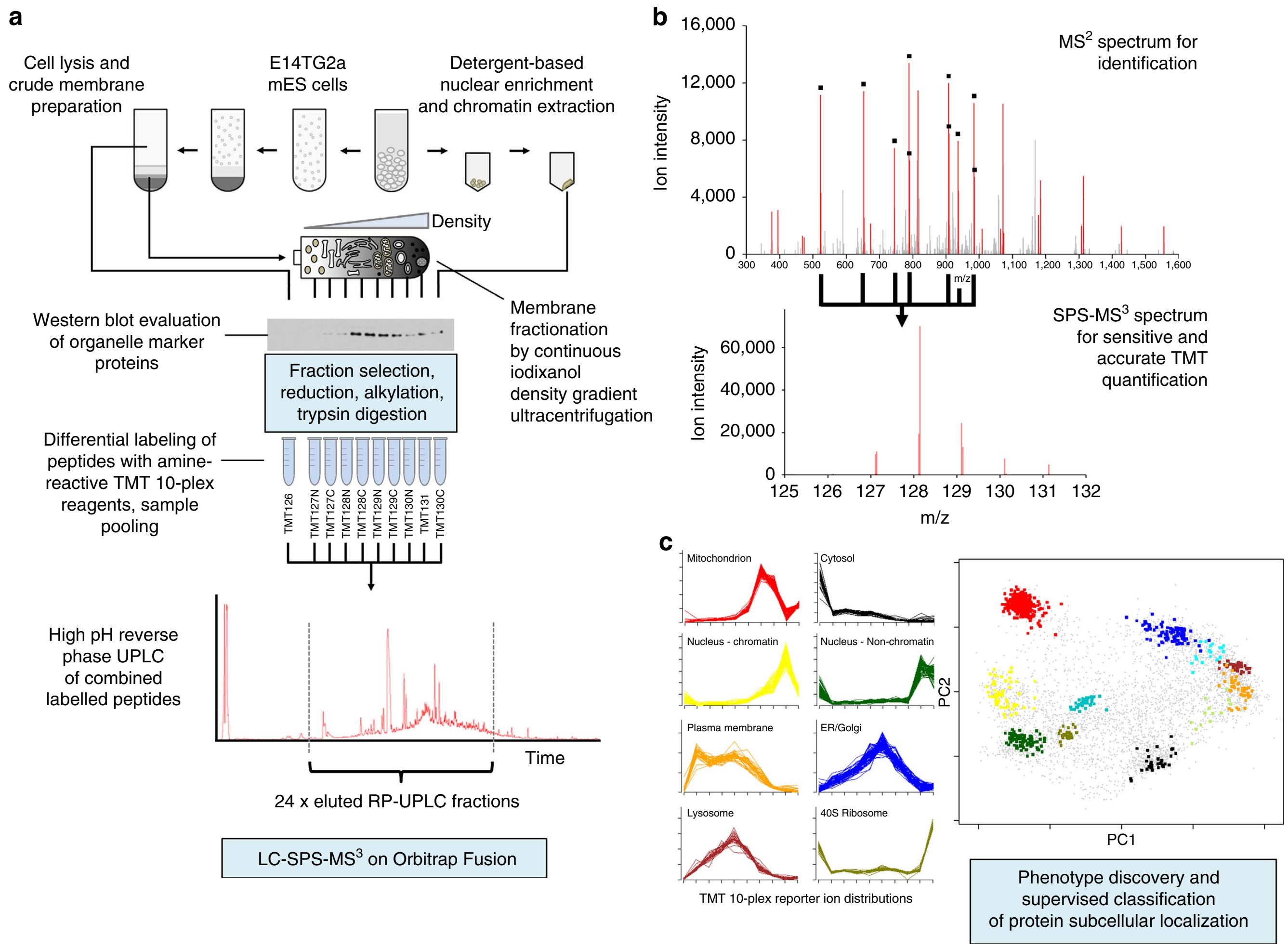

a
Cell lysis and crude membrane preparation
E14TG2a mES cells
Detergent-based nuclear enrichment and chromatin extraction
Density
Membrane fractionation by continuous iodixanol density gradient ultracentrifugation
Western blot evaluation of organelle marker proteins
Fraction selection, reduction, alkylation, trypsin digestion
Differential labeling of peptides with amine-reactive TMT 10-plex reagents, sample pooling
TMT126
TMT127N
TMT127C
TMT128N
TMT128C
TMT129N
TMT129C
TMT130N
TMT131
TMT130C
High pH reverse phase UPLC of combined labelled peptides
Time
24 x eluted RP-UPLC fractions
LC-SPS-MS^3 on Orbitrap Fusion
b
MS^2 spectrum for identification
Ion intensity
16,000
12,000
8,000
4,000
0
300 400 500 600 700 800 900 1,000 1,100 1,200 1,300 1,400 1,500 1,600
m/z
SPS-MS^3 spectrum for sensitive and accurate TMT quantification
Ion intensity
60,000
40,000
20,000
0
125 126 127 128 129 130 131 132
m/z
c
Mitochondrion
Cytosol
Nucleus - chromatin
Nucleus - Non-chromatin
Plasma membrane
ER/Golgi
Lysosome
40S Ribosome
TMT 10-plex reporter ion distributions
PC2
PC1
Phenotype discovery and supervised classification of protein subcellular localization

Figure 1 | Schematic of the hyperLOPIT workflow. (**a**) Ten enriched subcellular fractions were generated from E14TG2a murine ES (mES) cells. Peptides derived from each fraction were differentially labelled with TMT 10-plex reagents, and analysed by two-dimensional LC-SPS-MS^3. (**b**) SPS-MS^3 boosts the sensitivity of MS^3-based TMT quantification, while preserving the gains in quantitative performance relative to conventional MS^2, by selecting multiple peptide fragments rather than a single ion for MS^3 analysis. (**c**) TMT reporter ion distributions recapitulate the distribution of proteins across the fractionation scheme. Different organelles display characteristic distributions that may be used to determine their residents. The high dimensional data are presented in two-dimensions by PCA to provide an intuitive visualization of organelle separation.

SPS-MS^3 enhances quantitative performance. TMT quantification by conventional tandem mass spectrometry (MS^2) suffers from impaired quantitative accuracy and precision due to interference from contaminant peptides with similar chromatographic and mass-to-charge properties to the target peptide. Improved quantitative performance is achieved with an additional round of ion selection and fragmentation to purify the analyte from which TMT quantification is derived[17,18], but such MS^3-based methods result in substantially reduced sensitivity[19]. SPS-MS^3 balances the quantitative gains of MS^3 quantification with the sensitivity required for proteome-wide analysis[20,21]. Whereas in conventional MS^3 a single peptide fragment ion is selected for quantification, by using isolation waveforms with multiple frequency ‘notches’ to collect multiple peptide fragments, SPS-MS^3 improves ion statistics for quantification (Supplementary Fig. 1).

To assess the impact of SPS-MS^3 on quantitative performance, we evaluated the effect of the number of SPS-MS^3 ‘notches’ by comparing quantification derived from conventional MS^2, conventional MS^3 and SPS-MS^3 data acquisition. Adjusting the number of SPS-MS^3 notches altered the balance between quantitative performance and sensitivity. Increasing the number of notches augmented the TMT reporter ion signal intensity, and therefore the proportion of quantifiable spectra, but also reintroduces contaminant ions that distort quantification. Fewer notches result in lower reporter ion signal, with concomitant reduction in the number of quantifiable spectra. We found SPS-MS^3 with 10 precursors ‘notches’ to represent a suitable balance for global analysis, with 92.8% of the acquired SPS-MS^3 spectra yielding TMT reporter ion counts $>1 \times 10^5$—a comparable figure to that obtained with conventional MS^2 acquisition (Supplementary Table 1).

We then compared E14TG2a hyperLOPIT data sets acquired with conventional MS^2, and SPS-MS^3 with 10 notches. When comparing peptides derived from proteins with well curated localization, we observed SPS-MS^3 acquisition resulted in a significant improvement in quantitative accuracy over conventional MS^2 acquisition (Fig. 2 and Supplementary Figs 2 and 3). This gain in quantitative performance resulted in greater resolution of organelles from one another, and from the median of the total peptide population; tendency towards a unified distribution being characteristic of distorted TMT quantification (Supplementary Fig.18)[16,22].

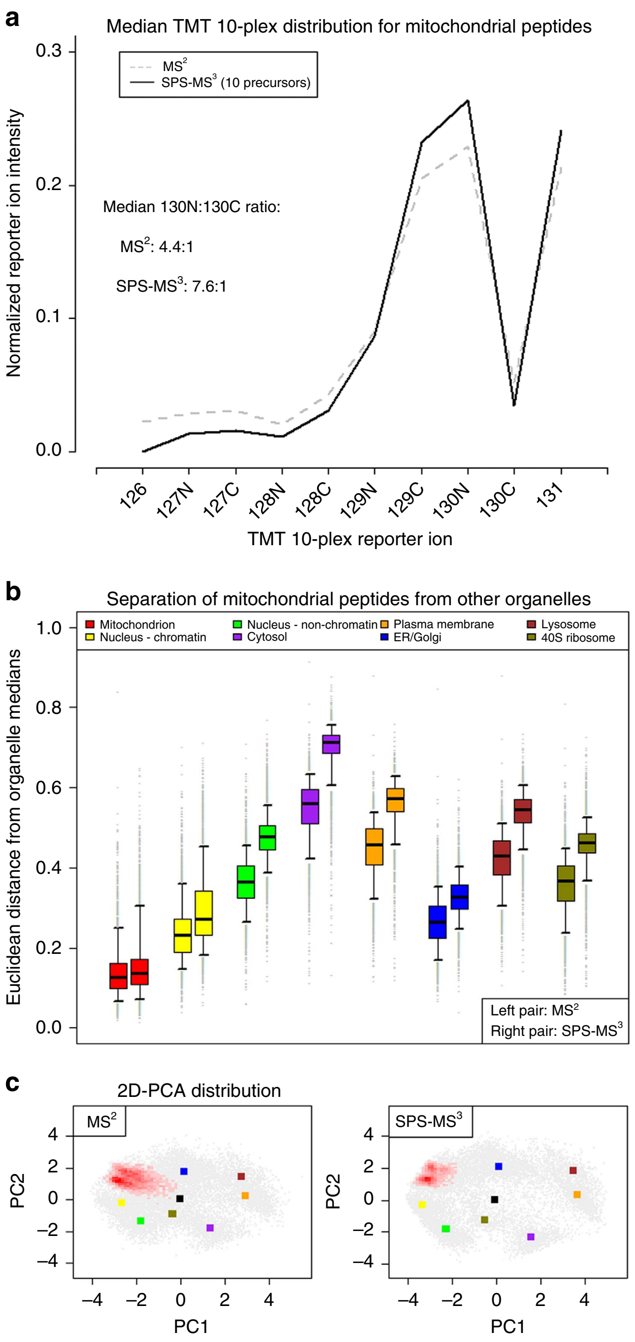


Figure 2 | SPS-MS^3 enhances resolution of the mitochondrion from other organelles. (**a**) SPS-MS^3 results in higher quantitative accuracy. The ratio between TMT channels expected to be enriched in mitochondria (129C and 130N) and channels that are expected to be depleted (for example, 130C) is markedly increased by SPS-MS^3. (**b**) The distance of mitochondrial PSMs from the median of eight organellar phenotypes. The distance of mitochondrial PSMs from the median mitochondrial distribution does not change significantly between MS^2 and SPS-MS^3, but the distances from other organelles are all significantly increased in SPS-MS^3 (Wilcoxon rank sum test *P* value $<2.2\times10^{-16}$), indicating greater organellar resolution. (**c**) When represented in two dimensions by PCA, the mitochondrial PSMs (red heat map) show less skew towards the origin, due to the improved specificity of SPS-MS^3 quantification. The median positions of other organelles are represented by coloured squares and individual proteins by grey points. Similar plots demonstrating the enhanced resolution of other organelles are presented in Supplementary Fig. 2.

While the increase in instrument duty cycle to perform SPS-MS^3 decreased the number of quantified peptide-spectrum matches (PSMs) from 137,912 to 61,090, the difference between the number of quantifiable protein groups was less substantial (7,114 versus 5,489)—a tolerable compromise given the gains in spatial resolution (Supplementary Fig.17).

HyperLOPIT provides an overview of protein localization. Over 6,000 protein groups were quantified in each of the three replicate experiments (Supplementary Data 1). The fractionation patterns observed in experiments 1 and 2 are highly consistent, while the resolution of secretory pathway organelles in experiment 3 was reduced owing to lower protein yields (Supplementary Fig. 4). Despite the lower resolution of some compartments in experiment 3, classification of proteins to subcellular compartments was highly reproducible, with <5% of proteins assigned contradictory localizations across the three experiments (Supplementary Fig. 5).

The intersect of experiments 1 and 2 (5,032 protein groups, Supplementary Data 1) was treated as a 20-plex data set for the analysis discussed in this article, which has previously been demonstrated to improve the attainable spatial resolution[23,24]. Experiment 3 was not included, as little additional resolution was obtained by further data fusion.

Using the pRoloc data analysis pipeline[25], an initial application of novelty detection[26] was conducted to identify and confirm the presence of organelle clusters in an unbiased data-specific manner, followed by supervised classification using a support vector machine (SVM) for final protein-organelle assignment. Applying SVM scoring thresholds based on concordance with gene ontology annotation, the steady-state localization of 2,855 out of a total of 5,032 protein groups were unambiguously determined (Fig. 3 and Supplementary Figs 6 and 7). Sub-nuclear resolution was also obtained, with proteins localized to chromatin, non-chromatin nuclear and nuclear lamina displaying distinct quantitative distributions. Manual curation of the data set also revealed co-localization of proteins for several other subcellular niches, including coatamer and clathrin-coated vesicles, and cytoskeletal fragments (Fig. 3).

Given the breadth of proteome coverage and subcellular resolution, hyperLOPIT can be used to investigate the organization of cells at multiple levels of scope. Insights into cell behaviour may be drawn from the localization of individual proteins, protein families and even functional networks, or by evaluating the protein content of particular subcellular compartments. In the following sections, we discuss how some of the layers of data can be interpreted, and the insights into the organization of pluripotent ES cells that may be drawn.

Organelle membership. The hyperLOPIT workflow enabled unambiguous assignment of 2,855 proteins to 14 discrete organellar and sub-organellar compartments in a single experiment. This number amounted to over 50% of the proteins identified, with the remainder displaying intermediate distributions as described in the Proteins in transit section. The classified proteins represent both residents of an organelle, as well as transient traffickers or cargo proteins that are captured in this position under these experimental conditions and in this particular cell type. The catalogue of organelle members is therefore context specific and must be viewed as a subcellular snapshot of a dynamic system.

Approximately 83% of proteins classified by hyperLOPIT carry localization-specific gene ontology annotation, with only 39% of these proteins annotated based on direct assay evidence. This data set therefore provides new experimental

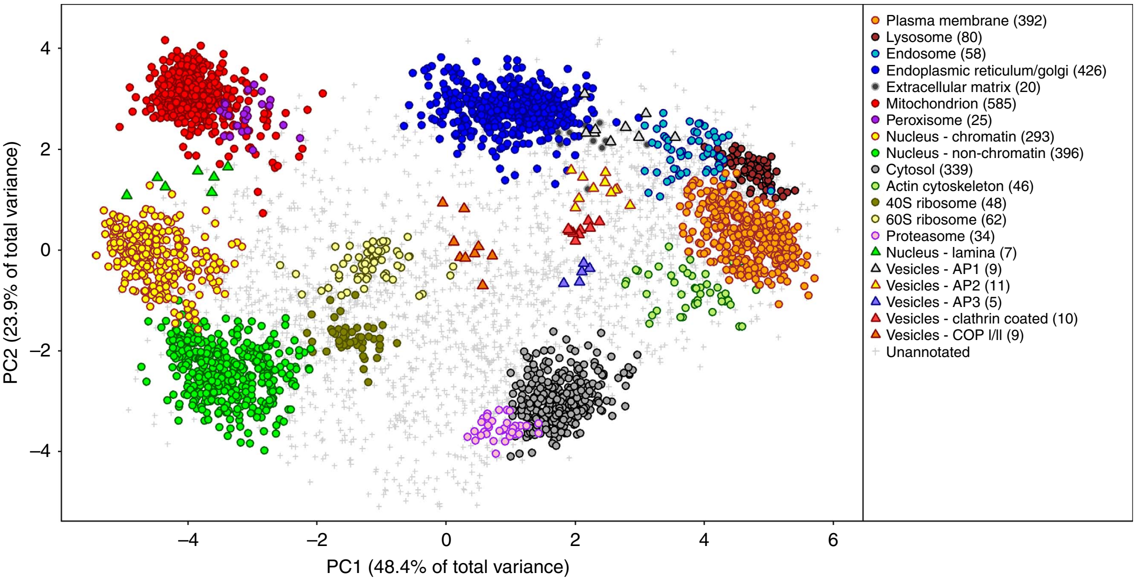


Figure 3 | PCA representation of E14TG2a mouse ES cell hyperLOPIT data. Each data point represents a protein group, and proteins with correlated TMT reporter ion distributions cluster together in PCA space. Coloured circles represent subcellular compartments that have been classified by SVM and coloured triangles represent smaller organellar phenotypes that were manually curated. While two dimensions describe much of the spatial resolution in these data, separation of some compartments, such as the mitochondrion and peroxisome, is only apparent in the lower principal components (Supplementary Fig. 5).

evidence for the localization of 1,775 murine proteins, ∼350 of which currently lack organelle-specific localization information in the UniProt database (Supplementary Data 1). It should be noted that these values include proteins that are annotated as cytoplasmic. Since the term cytoplasm incorporates the cytosol, cytoskeleton, organelles and other cellular features other than the nucleus and cell surface, such annotation lacks the specificity to be suitably informative in this context. HyperLOPIT is therefore a useful tool for supplementing database annotation of protein localization with high throughput. Another key strength of the approach is that protein localization may be determined in particular cell type(s) and cell state(s) of interest—information that is typically lacking in gene ontology annotation or computational prediction of localization.

We compared the protein localizations determined by hyperLOPIT with another recently published plasma membrane proteome for the same cell line. Bausch-Fluck *et al.*[27] identified plasma membrane proteins by cell surface biotinylation and affinity purification-mass spectrometry. In this study the authors categorized their cell capture data into three groups: High confidence, for proteins with UniProt keywords including ‘Cell junction’, ‘Cell membrane’ and ‘Secreted’; putative, for proteins with predicted transmembrane domains, but none of the above keywords assigned; and non-specific, for other identified proteins, which were assumed to be abundant contaminant proteins. Proteins labelled as high confidence in this study corroborate with their localization in the hyperLOPIT data, as almost all such proteins localize to the plasma membrane or endosomes—suggesting extensive recycling of some of these surface proteins. HyperLOPIT also confirms plasma membrane localization for many proteins labelled as putative in the Bausch-Fluck study (Supplementary Fig. 8). The third category, non-specific proteins, consists mostly of proteins found in non-surface localizations by hyperLOPIT. The strong correlation of results demonstrates that hyperLOPIT may also be used as an orthogonal validation method for targeted localization studies.

Proteins in transit. Not all proteins are expected to partition into discrete subcellular compartments. While over half of the quantified proteins were classifiable to a single and unambiguous location, many proteins were found to have distributions that did not closely correlate with those of the 14 classified subcellular compartments. These less discrete distribution patterns occur for several reasons.

First, proteins present in multiple compartments will adopt quantitative distributions that reflect their steady-state subcellular enrichment. β-catenin for example, has been robustly classified to the plasma membrane despite the fact that this protein is known to localize to the adherens junctions, cytosol and nucleus. Classification to the plasma membrane reflects its relative enrichment at the adherens junction in ES cells cultured under self-renewing conditions, which has been previously demonstrated by immunofluorescence microscopy[28].

There are cases where the steady-state localization is not so skewed towards one of multiple compartments. For example, many components of the nuclear import and export machinery display distributions consistent with mixed localization between the cytosol and nucleus (Supplementary Fig. 9). These proteins were not classified as unambiguous residents of either the nucleus or cytosol; their distribution patterns accurately reflect their true subcellular localization. Other proteins displaying mixed localizations include signalling cascade effectors, such as Erk-2 and the adaptor protein Grb-2, components of the MAP kinase signalling pathway that distribute between the plasma membrane and cytosol, reflecting the dynamic transitions between the two locations. For some proteins with mixed localization the distribution patterns are quite complex. Mcl-1, a Bcl-2 family member protein, displays a three-way mixed localization between the endoplasmic reticulum, mitochondrion and nucleus, and therefore falls between the three organelle clusters in principal component analysis (PCA) space; this broad distribution is consistent with previous confocal microscopy analysis of its human orthologue[29].

The steady-state distribution of transitory proteins can provide information about the state of the cell population. For example, Tfe3, a helix–loop–helix family transcription factor and modulator of the exit of ES cells from pluripotency, was observed with mixed localization between the cytoplasm and nucleus (Supplementary Fig. 9), consistent with immunocytochemistry data in ES cells. In the pluripotent state, Tfe3 localizes to both the nucleus and cytosol, and regulates the expression of key pluripotency factor Esrrb. In the early stages of differentiation, it is excluded from the nucleus[30].

Second, intermediate subcellular distributions may represent proteins residing in organelles other than those used in the SVM classification. These compartments typically have their own physicochemical properties, and their resident proteins therefore co-localize, but are not sufficiently enriched for their distribution patterns to be fully resolved from those of other subcellular compartments. For example, known components of the clathrin-AP3 trafficking vesicles were found to co-localize away from other organelles and vesicles, but were not distinct from other proteins with mixed localization. Some proteins with correlated distributions to the clathrin-AP3 vesicles are plausible vesicular components, such as syntaxin-18, whereas others are transitory cytosolic and cytoskeletal proteins such as serine/threonine kinase Pak4.

Finally, the observed localization patterns may represent proteins that comprise or are tethered to the cytoskeleton. The fractionation pattern of the cytoskeleton is not easily predicted given its broad interconnectivity with other subcellular components. We observed groups of cytoskeletal proteins with a variety of distinct distributions, including actin, actin-capping modulators, microtubules, microtubule organizing centre components, dyneins, kinesins and myosins (Supplementary Fig. 10). While we did not perform classification on these smaller phenotypes, it is possible that uncharacterized proteins that co-localize with these niches are cytoskeletal components.

Organelle structure. The ability of hyperLOPIT to gain insight into very dynamic processes within the cell is exemplified by the Golgi apparatus. The position of well documented Golgi apparatus marker proteins reveals an unexpected observation about these cells. As can be seen from Fig. 4a, these markers fall into four categories; proteins that co-cluster with endoplasmic reticulum (ER) markers, proteins that lie between the ER/Golgi and endosomal compartments, proteins that cluster with plasma membrane markers, and proteins whose steady-state location does not correlate with any of the 14 classified compartments. The proteins that cluster along with ER markers are generally annotated as being *cis*-Golgi proteins, including alpha-mannosidase 2 (MA2A1)[31] and members of the SNAP receptor complex

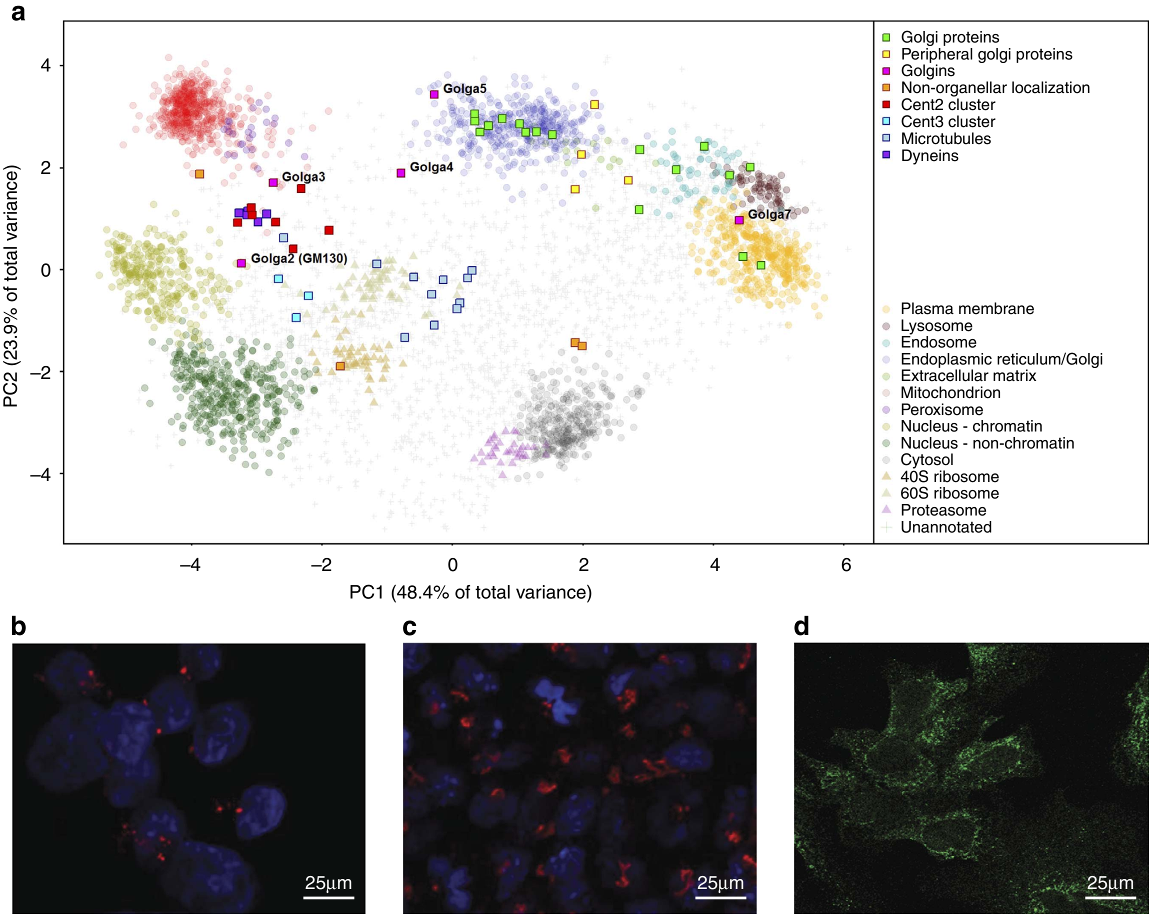


Figure 4 | Steady-state location of Golgi apparatus marker proteins. (**a**) PCA plot of data with Golgi marker proteins overlaid Golgi proteins (green) peripheral Golgi proteins (yellow), Golgins (pink), microtubular proteins (light blue), dyneins (purple), centrin 2 and associated proteins (red) centrin 3 and associated proteins (mid-blue), Golgi proteins with aberrant localization (orange). (**b**) Immunocytochemistry using anti-GM130 (red) on a self-renewing population of E14TG2a cultured in serum and LIF showing a punctate localization not consistent with the Golgi Apparatus. (**c**) Immunocytochemistry using anti-GM130 on a population of E14TG2a after 4 days in N2B27, a media promoting neural differentiation. (**d**) Immunocytochemistry on a self-renewing population of E14TG2a cultured in serum and LIF using anti-KDEL (green; ER lumenal retention signal) to verify typical endoplasmic reticular structures encapsulating the nucleus.

(GOSR2)[32]. Proteins that cluster towards the endosomal markers are associated with the *trans*-Golgi, such as the copper-transporting ATPase (ATP7A)[33] and vacuolar protein sorting-associated protein 45 (VSP45)[34]. Several proteins form part of the plasma membrane cluster including Ras-related protein Rab-6A (RAB6A)[35] and Golgin 7 (GOGA7), which are thought to be involved in protein transport from the Golgi to the cell surface[36]. In general those proteins known to be peripherally associated with Golgi membranes lie in an intermediate position on the boundary of the ER/Golgi cluster, for example, the Golgi phosphoprotein GOLP3 and oxysterol-binding protein 1.

It is curious to note that some Golgi marker proteins occupy a steady-state position far removed from the ER/Golgi or endosomal vesicles. GM130 (GolginA2) is a *cis*-Golgi matrix protein, and is thought to be involved in regulation of centrosomes during interphase[37]. On the PCA plot (Fig. 4A), GM130 is intriguingly at a steady-state position between centriolar proteins such as pericentriolar material 1 protein (PCM1), centrin 2 (CETN2), spindle and centriole-associated protein (SPICE), pericentrin (PCNT) and another cluster of proteins containing gamma tubulin subunits and centrin 3 (CETN3)[38–40]. This is consistent with the role of GM130 to act as a tether between the Golgi and centrosomes in interphase[41]. To confirm our observations, we performed immunocytochemistry imaging with an anti-GM130 antibody, which supported the theory that this protein may be associated with the centriole in pluripotent mouse ES cells (mES; Fig. 4b). Interestingly, we noted that GM130 instead appeared to be associated with Golgi-like structures when the cells are transferred to a media promoting neural differentiation (N2B27 medium; Fig. 4c,d). The reasons behind the difference in distribution of GM130 observed between self-renewing and differentiating ES cells may reflect the relative dwell times of the two sets of cells in the stages of the cell cycle[42], and it has previously been reported that mES cells have a truncated G1 phase[43]. The distribution of GM130 and is proximity to centriolar and pericentriolar matrix proteins may reflect that the majority of pluripotent cells are in the later stages of the cell cycle. Although the data do not show direct tethering of Golgi and centrosomal complex structures, this is just one example of how these data provide a snapshot of cellular behaviour.

Protein complexes. In addition to organelles, we observed that many protein complexes display highly correlated distribution patterns. Co-localization of highly abundant macromolecular protein complexes, such as ribosomal subunits and the mitochondrial ATP synthase complex by LOPIT has been previously demonstrated[44]. The depth of proteome coverage in this study, coupled with the high subcellular resolution derived from precise quantification, enabled many more complexes to be detected. We selected 30 examples of highly curated protein complexes listed in KEGG[45] and Reactome[46,47], and annotated their distributions in the hyperLOPIT data (Fig. 5). There are many other protein complexes represented in the data set that we have not curated, but which readers of this study can explore according to their particular interest using pRolocGUI.

Proteins with multiple functional roles within the cell may be involved in more than one protein complex. Neither the extent of this, nor the distribution of components between these complexes is typically captured in large-scale complex purification studies. Inspection of the steady-state position of complex components within hyperLOPIT data indicates that components have differing positions from the core complex members, and may even indicate those that function in a regulatory manner. For example, the steady-state location of the TFIID complex is nuclear in these data, however, Taf7 has a steady-state distribution that is distinct from the other complex components

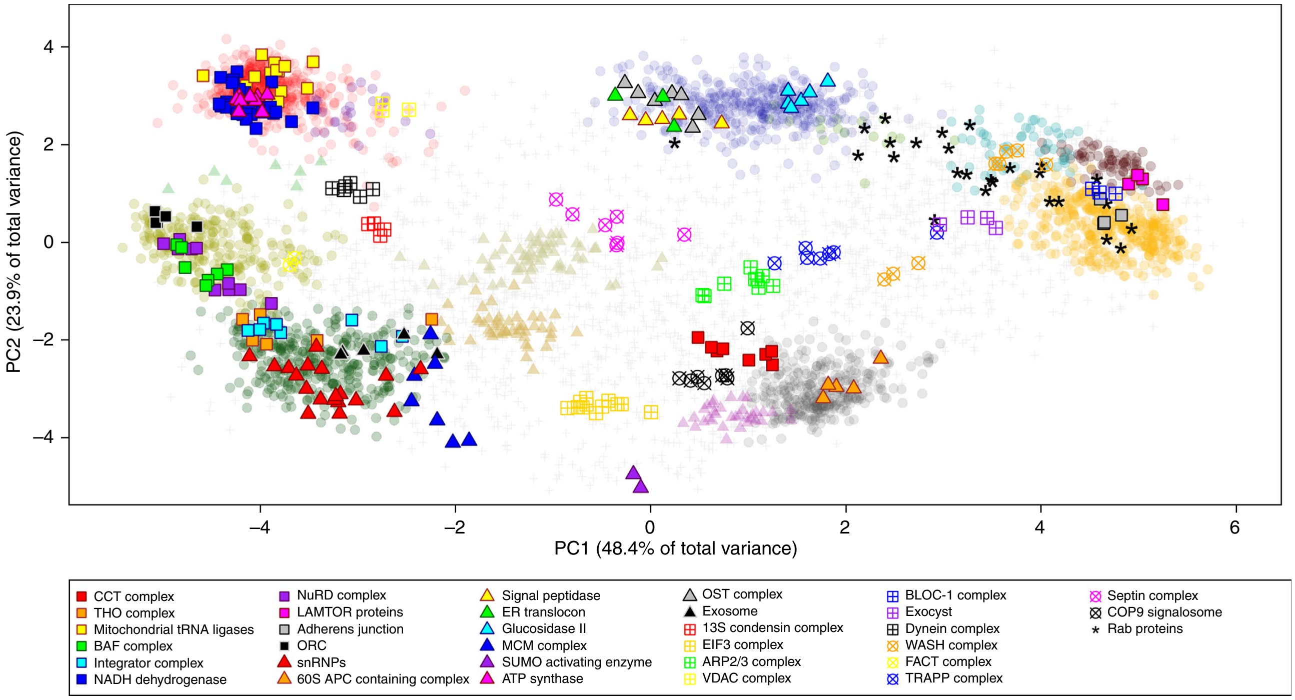


Figure 5 | Examples of soluble and organellar protein complexes. Protein complexes are distributed throughout principal component space. Some soluble protein complexes display characteristic distributions that may reflect mixed localization, partitioning of the cytoskeleton by the fractionation scheme, or the unique sedimentation properties of these macromolecular structures. Sub-organellar distribution of complexes were observed in both ER, mitochondrion and nucleus.

(Supplementary Fig. 9). This may reflect the fact that Taf7 is thought to dissociate from the pre-initiation TFIID complex following initiation of transcription[48]. Another example is the exocyst complex, where Exoc8 is localized away from the core exocyst complex in these data, and co-distributes with its known binding partners Par6 and RalA[49].

To demonstrate the utility of these data in evaluating other high-throughput data sets, we annotated the murine orthologues of proteins that were characterized in a census of human soluble protein complexes[50]. By adding this subcellular context to the analysis, we see that well-established components of protein complexes tend to co-localize, whereas novel assignments and putative uncharacterized complexes display more varied localization (Supplementary Fig. 11). Novel assignments that co-localize with the 'core' complex might be assumed to be stable interactors, whereas those with different subcellular distributions may be transient interactors or false assignments. For example, two of eight novel components of the 39S mitochondrial ribosomal subunits assigned by Havugimana *et al.*[50] were found to co-localize to the mitochondrion, whereas the remaining six were distributed in the nucleus, cytosol and secretory pathway (Supplementary Fig. 11C). The two co-localized proteins (Ict1 and Mrp63) both feature mitochondrial signal peptides and have molecular functions consistent with translation. The mouse and human variants of the six other putative interactors do not contain a distinct signal peptide, and are therefore probable false assignments in the census data set.

Functional networks. The high proteome coverage generated by hyperLOPIT enabled us to determine the localization of many proteins associated with pluripotency and differentiation, including components of the core transcriptional network of pluripotent cells such as Sox2, Oct4 and Nanog. Also, impressively, the subcellular distributions of the components of FGF/MAPK, canonical Wnt, Notch, BMP/SMAD, Nodal, Ras and Hippo signalling pathways are apparent, paving the way for using hyperLOPIT to determine modulation in the location of signalling proteins and their effectors on activation/deactivation of multiple signalling pathways during, for example, differentiation (Supplementary Fig. 12). During self-renewal, Wnt/β-catenin and FGF/MAPK signalling are maintained at a low level[28]. Consistent with this, we observe that β-catenin is firmly located in the plasma membrane and that Sprouty proteins, negative regulators of FGF/ERK signalling, are also associated with the membrane, where they act to inhibit the early stages of FGF signalling. We also observe extranuclear localization of Smad2/5, consistent with the known low Nodal/Activin signalling in self-renewal[42]. To our knowledge this is the first instance in which it is possible to have a snapshot of all signalling pathways in one cell.

To evaluate the spatial distribution of interaction partners for the core pluripotency transcription factor triad of Oct4, Sox2 and Nanog, we overlaid information from three high-throughput protein–protein interaction studies onto the E14TG2a hyperLOPIT data set[51–53]. As might be expected, the seven interaction partners that were common to all three bait proteins displayed steady-state enrichment to chromatin (Supplementary Fig. 13D). Three of these common interaction partners (Chd4, Mta1 and Mta2) are components of the NuRD complex, demonstrating the importance of chromatin remodelling for maintenance and exit from the pluripotent state[54]. While the interactome was found to be chromatin centric, each of the transcription factors was also reported to interact with proteins that we observed with subcellular enrichment away from chromatin by hyperLOPIT (Supplementary Fig. 13A–C). The functional implications of these non-chromatin interactors are of potential interest, as their differential localization suggests that trafficking of the interaction partners into the nucleus, or of the transcription factor to extranuclear locations, plays a role in modulating the interaction. Alternatively, protein–protein interactions that are far removed may be indicative of a false discovery, as described previously for the 39S ribosomal subunit.

Protein isoforms. HyperLOPIT can also provide information on the localization of protein isoforms. The impact of post-transcriptional modification on protein location has been previously documented[55], however observation of the differences in subcellular location of closely related isoforms has not been previously possible using LOPIT, as the accuracy of quantification was insufficient to reliably characterize the few peptides, or often single peptide, which distinguish protein isoforms. With the enhanced quantitative performance of TMT quantification by SPS-MS[3], measurements based on few peptides still yield reliable reporter ion distributions that allow us to characterize localization. We observed unique evidence for 25 and 26 pairs of protein isoforms (distinct proteins sharing the same gene name) in experiments 1 and 2, respectively (Supplementary Data 1). Some of the detected protein isoforms were found to co-localize, while others were found to be differentially localized.

For example, two murine isoforms of Leucine aminopeptidase 3 (Lap3) have been reported to arise from alternative translational initiation codons in the same mRNA; a 'long' canonical isoform, and a 'short' isoform with a 31 residue N-terminal truncation. We observed the 'long' isoform of Lap3 with unambiguous mitochondrial localization, whereas the 'short' isoform was found with a steady-state distribution between the cytosol and plasma membrane (Supplementary Fig. 14A). Protein-localization algorithms based on primary sequence, such as WoLF PSORT (ref. 56), support the observation that the long isoform of Lap3 localizes to mitochondria, whereas the short isoform does not (although the specific localization of the short isoform within the cytoplasm is not predictable from sequence alone). This suggests that the two isoforms, while sharing a common catalytic activity, fulfil separate biological roles due to differential localization. The differential localization is achieved through alternative translation that incorporates or excludes an N-terminal mitochondrial target signal. Further credence is given to the dual localization determined by hyperLOPIT by the fact that predicted functional partners of Lap3 (Anpep, Cat, Ggt1, Gss, Hspd1 and Pycr), as reported by the STRING protein interaction database (v9.1 (ref. 57)), are found to localize to the mitochondrion, plasma membrane, and cytosol—the three subcellular compartments described by the steady-state localization of the two Lap3 isoforms.

The differential localization of some other isoforms is not as straightforward to interpret. Two isoforms of Dnmt1, a DNA (cytosine-5)-methyltransferase that modulates ES cell pluripotency by maintaining CpG methylation patterns[58] were identified; a canonical 'long' isoform, and 'short' isoform with an 118 residue N-terminal truncation that removes the Dmap interacting domain. The canonical isoform distribution was consistent with localization to chromatin, and closely co-localizes with Dmap1. The short isoform also appears to localize to the nucleus, but with an atypical distribution profile that while most similar to chromatin, is distinct from the typical chromatin profile (Supplementary Fig. 14B). The significance of this differential nuclear distribution is not immediately clear[59].

While the number of isoforms pairs identified here is relatively modest, the depth of attainable proteome coverage will increase as the speed and sensitivity of high-resolution accurate mass spectrometry continues to advance, and will allow more of these isoforms to be detected. Accurate quantification of proteins based on a single peptide measurement will also permit characterization of post-translationally modified peptides. The ability to functionally characterize post-translational variants of proteins is an exciting avenue of research, as such variants introduce an additional dimension of biochemical complexity that is not easily evaluated with high throughput.

Discussion

The hyperLOPIT technique is a systematic and accurate assay for characterizing the localization of thousands of proteins in a single experiment, and is applicable to many biological model systems. Application of this technology to a self-renewing population of ES cells has generated the most extensive analysis of protein localization in a stem cell line to date. The data contain information about organelle residency of proteins, sub-organellar structure, the impact of isoform status on location and dynamic localization of proteins. This creates a reference for mapping proteins relative to each other in pluripotency and differentiation.

HyperLOPIT data may also be used to provide a spatial context for pre-established protein complexes and functional networks. For example, our results provide molecular support for some functional observations about the state of several signalling pathways in ES cells. HyperLOPIT can also be used as an orthogonal approach for validation of targeted localization studies, as we have demonstrated with the cell surface proteome generated by Bausch-Fluck *et al.*[27]

The E14TG2a data set we present acts both as a resource for interrogating the subcellular location of proteins of interest to researchers, and also acts as a scaffold onto which other high content data sets may be mapped to assist in their interpretation. HyperLOPIT is a powerful tool for gaining insights into fundamental post-translational processes governing stem cell behaviour.

Methods

Cell culture. Murine pluripotent ES cells (cell line E14TG2a) a kind gift from Professor Austin Smith, University of Cambridge, and available from American Type Culture Collection (CRL-1821), were maintained in culture on gelatinized flasks in a media containing fetal bovine serum supplemented with leukaemia inhibitory factor (LIF), as previously described[60]. Approximately 10^8 cells were collected by trypsinization and washed several times with phosphate buffered saline. Cell suspension (10%) was aliquoted for chromatin extraction, while the remaining 90% was used for membrane fractionation.

Immunocytochemistry. E14TG2A mES cells were plated and stained as described in ref. 42 using anti-GM130 (AbCam—EP829Y) and KDel (AbCam—ab50601), and imaged by confocal microscopy.

Cell lysis and subcellular fractionation. For density-gradient ultracentrifugation, cell pellets were resuspended in 15 ml lysis buffer (0.25 M sucrose, 10 mM HEPES pH 7.4, 2 mM EDTA, 2 mM magnesium acetate) containing protease inhibitors (Roche), and lysed with a ball-bearing homogenizer (Isobiotec) on ice. Lysate viscosity was reduced by treatment with 25 U ml^{-1} benzonase endonuclease (Invitrogen) for 20 min at room temperature. Insoluble cellular debris was removed by centrifugation at 200*g*, 5 min at 4 °C. The supernatant was retained and the centrifugation step was repeated a further two times.

Optiprep density-gradient medium (60% w/v iodixanol, Sigma) was diluted to a working solution of 50% w/v iodixanol with 6 × lysis buffer (60 mM HEPES pH 7.4, 12 mM EDTA pH 8.0, 12 mM magnesium acetate), containing protease inhibitors. Solutions with varying concentrations of iodixanol were then made by mixing the iodixanol working solution and lysis buffer. The iodixanol concentration of each solution was verified by measuring refractive index using a handheld refractometer (Reichert Technologies).

The lysate was divided between several 5 ml polyallomer ultracentrifuge tubes (Beckman), and underlaid with 0.8 ml of 6% w/v iodixanol solution, and then with 0.8 ml of 25% w/v iodixanol solution. Samples were centrifuged in an Optima XL-80 ultracentrifuge (Beckman), SW55Ti rotor at 100,000*g*, 60 min at 4 °C. The resulting supernatant was retained as a 'soluble fraction' (enriched in cytosolic proteins), and crude membranes were collected from the interface of the two iodixanol layers. The crude membrane fraction was diluted with lysis buffer and pelleted by ultracentrifugation in the SW55Ti rotor at 200,000*g*, 40 min at 4 °C to remove any residual cytosol from the membranes. The washed membrane pellet was then resuspended in 25% w/v iodixanol solution, and underlaid beneath a pre-formed gradient composed of 8, 12, 16 and 20% iodixanol layers, which was left for 8 h at 4 °C to diffuse to linearity. The continuous density gradient was centrifuged at 100,000*g* for 8 h in a VTi65.1 rotor at 4 °C with slow braking to minimize gradient disruption. Following ultracentrifugation, 20 × 0.5 ml gradient fractions were collected using an Auto Densi-Flow peristaltic pump with meniscus tracking probe (Labconco). The refractive indices of all fractions were measured to determine the shape of the final gradient. Each fraction was then diluted with 0.8 ml lysis buffer, and centrifuged in a TLA-55 fixed angle rotor at 180,000*g* in an Optima MAX-XP benchtop ultracentrifuge (Beckman), 20 min at 4 °C. This centrifugation step was repeated for all fractions, the supernatant was discarded and the resulting membrane-enriched pellets were stored at − 20 °C. Four volumes of chilled acetone were added to the cytosolic fraction and the chromatin-enriched fraction, and protein precipitation was carried out overnight at − 20 °C. The acetone samples were centrifuged and air-dried before solubilization in 8 M urea, 0.1% SDS, 25 mM tetraethylammonium bromide (TEAB; pH 8.5). The density-gradient membrane-enriched pellets were also resuspended in 8 M urea buffer. Samples were briefly sonicated on ice to ensure re-solubilization.

The reproducibility of gradients from three independent biological replicates can be found in Supplementary Table 2.

Chromatin extraction and enrichment. Chromatin extracts were prepared as previously described[61]. Briefly, cells were resuspended in chromatin buffer A (10 mM HEPES pH 7.9, 10 mM KCL, 1.5 mM $MgCl_2$, 0.34 M sucrose, 10% glycerol, 1 mM dithiothreitol) with protease inhibitors. Triton X-100 was added to a concentration of 0.1% v/v and incubated on ice for 8 min to lyse the cells. Nuclei were pelleted by centrifugation at 1,300*g*, 5 min at 4 °C. The nuclear pellet was resuspended in chromatin buffer B (3 mM EDTA, 0.2 mM EGTA, 1 mM dithiothreitol) with protease inhibitors and incubated for 30 min on ice. Samples were then centrifuged at 1,700*g*, 5 min at 4 °C. The chromatin-enriched pellet was washed in chromatin buffer B, re-pelleted and stored at − 20 °C.

Protein digestion and TMT 10-plex labelling. Protein concentrations were determined by BCA assay (Thermo Fisher Scientific) as per the manufacturer's instructions. Optimal fractions were selected for TMT 10-plex labelling based on western blot evaluation of organelle marker proteins, protein concentration and refractive indices of membrane fractions. Protein (50 μg) from 10 differentially enriched subcellular fractions (8 membrane fractions from the density gradient, plus cytosol and chromatin-enriched fractions, Supplementary Table 3) was reduced, alkylated and digested with trypsin. Briefly, each sample was made up to a total volume of 50 μl with 25 mM TEAB and reduced and alkylated. Disulfide bonds were reduced with 5 μl of 200 mM tris(2-carboxyethyl)phosphine, 1 h at 37 °C, followed by alkylation of cysteine residues with 5 μl of 375 mM iodoacetamide, 30 min at room temperature. Samples were then diluted tenfold with 25 mM TEAB and digested with sequencing grade trypsin (Promega) for 1 h with a 1:40 enzyme:protein ratio, 37 °C. An additional aliquot of trypsin at 1:40 concentration was added and incubated overnight at 37 °C. Trypsin digests were centrifuged for 10 min at 13,000*g* to remove any insoluble matter, then reduced to dryness by vacuum centrifugation.

While the TMT tags were equilibrating to room temperature, peptide samples were resuspended in 30 μl 1 M TEAB and 70 μl isopropanol. The solubilized samples were transferred into the tag vials and placed on a shaker for 2 h at room temperature. The reaction was quenched by addition of 8 μl 5% hydroxylamine for 30 min. The labelled samples were then combined and reduced to dryness by vacuum centrifugation. C18 solid-phase extraction was performed using Sep-Pak cartridges (100 mg bed volume, Waters) and peptides were eluted in 70% acetonitrile + 0.05% acetic acid. The eluate was again reduced to dryness by vacuum centrifugation, and resuspended in 20 mM ammonium formate (pH 10.0), for high pH reversed-phase liquid chromatography.

Sample fractionation. Desalted peptides were resuspended in 0.1 ml 20 mM ammonium formate (pH 10.0) + 4% (v/v) acetonitrile. Peptides were loaded onto an Acquity bridged ethyl hybrid C18 UPLC column (Waters; 2.1 mm inner diameter × 150 mm, 1.7 μm particle size), and profiled with a linear gradient of 5–60% acetonitrile + 20 mM ammonium formate (pH 10.0) over 60 min, at a flow rate of 0.25 ml min^{-1}. Chromatographic performance was monitored by sampling eluate with a diode array detector (Acquity UPLC, Waters) scanning between wavelengths of 200 and 400 nm. Fractions were collected at 1 min intervals. Twenty-four fractions representing peak peptide elution were selected for mass spectrometry analysis and resuspended in 0.05% trifluoroacetic acid. Approximately 1 μg peptides were loaded per liquid chromatography-mass spectrometry run.

Mass spectrometry. All mass spectrometry experiments were performed on an Orbitrap Fusion coupled with a Proxeon EASY-nLC 1000 (Thermo Fisher Scientific). Peptides were separated on a Proxeon EASY-Spray column (Thermo

Scientific; 50 cm × 75 μm inner diameter, 2 μm particle size and 100 Å pore size). Separation was achieved by applying a 5–25% gradient of acetonitrile + 0.1% formic acid over 95 min at 300 nl min^{-1}, followed by 25–40% acetonitrile + 0.1% formic acid over 10 min. An electrospray voltage of 1.8 kV was applied to the eluent via the EASY-Spray column electrode.

The Orbitrap Fusion was operated in positive ion data-dependent mode for both MS^2 and SPS-MS^3 methods. For the MS^2 method, the full scan was performed in the Orbitrap in the range of 300–1,600 m/z at nominal resolution of 1.2×10^5, followed by selection of the most intense ions above an intensity threshold of 2×10^4 for high-energy collisional dissociation (HCD)-MS^2 fragmentation. Ion filtering for MS^2 events was performed by the quadrupole with a transmission window of 1.5 m/z. HCD fragmentation was performed with 40% normalized collision energy, followed by analysis of fragment ions in the Orbitrap with nominal resolution of 6×10^4. The number of HCD-MS^2 events between full scans was determined on-the-fly so that the duty cycle was fixed at 3 s.

The automatic gain control (AGC) settings were 4×10^5 ions and 1×10^5 ions, and maximum ion accumulation times to 50 and 120 ms, for full and MS^2 scans, respectively. Ions with 1+ or undetermined charge state were excluded from MS^2 selection. Ions within a ± 10 p.p.m. m/z window around ions selected for MS^2 were excluded from further selection for fragmentation for 35 s.

For the SPS-MS^3 method, the full scan parameters were identical to those for the MS^2 method. The most intense ions above a threshold of 2×10^4 were selected for collision induced dissociation (CID)-MS^2 fragmentation, with an AGC target and maximum accumulation time of 1×10^4 and 70 ms. Mass filtering was performed by the quadrupole with 1.5 m/z transmission window, followed by CID fragmentation in the linear ion trap with 35% normalized collision energy. SPS was applied to co-select 10 fragment ions for HCD-MS^3 analysis. SPS ions were all selected within the 400–1,000 m/z range, and were set to preclude selection of the precursor ion and TMTC ion series[62]. AGC targets and maximum accumulation times were set to 1×10^5 and 120 ms. Co-selected precursors for SPS-MS^3 underwent HCD fragmentation with 55% normalized collision energy, and were analysed in the Orbitrap with nominal resolution of 6×10^4. The number of SPS-MS^3 spectra acquired between full scans was restricted to a duty cycle of 3 s.

To assess the effect of using different numbers of precursors for SPS, ions were selected from full scans as described above. For each selected peptide ion, a sequence of six spectra was generated (conventional MS^2, SPS with 15, 10, 5 and 2 precursors, and conventional MS^3). The precursor ion for conventional MS^3 was selected as the most intense ion within the 400–950 m/z range, excluding the unfragmented peptide and TMTC ion series, and isolated with a 2 m/z selection window. Conventional MS^2 and SPS-MS^3 were performed with the parameters described previously. The duty cycle for the sequence of scans was fixed at 6 s, with each sequence of six scans taking ~1.5–2.5 s.

Data processing. Raw data files were processed using Proteome Discoverer (v1.4, Thermo Fisher Scientific), interfaced with Mascot server (v2.3.02, Matrix Science). Mascot searches were performed against SwissProt mouse database (March 2013, 24,481 sequences), with carbamidomethylation of cysteine, and TMT 10-plex modification of lysine and peptide N termini set as modifications. For the MS^2 method, in which identification was performed at high resolution in the Orbitrap, precursor and fragment ion tolerances of ± 20 p.p.m. and ± 0.2 Da were applied. For the SPS-MS^3 method, in which identification was performed at lower resolution in the linear ion trap, tolerances of ± 20 p.p.m. and ± 0.5 Da were applied. Up to two missed tryptic cleavages were permitted. Searches were also performed against a sequence scrambled 'decoy' database. The PSMs for the 'forward' and 'decoy' searches by Mascot were re-scored using the Percolator algorithm to yield a more robust false discovery rate[63].

TMT 10-plex quantification was also performed by Proteome Discoverer by calculating the sum of centroided ions within ± 2 mmu window around the expected m/z for each of the 10 TMT reporter ions. For SPS-MS^3 methods, quantification was performed at the MS^3 level. Spectra with more than four missing reporter ion values were excluded from quantification, and remaining missing values were set as zero for downstream analysis. For protein-level reporting, protein grouping was enabled, and values were calculated from the median of all quantifiable PSMs for each group. TMT values were then reported as a ratio to the sum of reporters in each spectrum (that is, the sum of the 10 values for each spectrum was equal to 1).

To evaluate the effect of the number of SPS notches on quantitative performance, data were first processed as described above. SPS data were then extracted from scan headers in the raw data files by a VB.NET script using MSFileReader libraries (Thermo Fisher Scientific). SPS data and reporter ion quantification were then paired with the peptide identification information reported by Proteome Discoverer.

Machine learning and multivariate data analysis. The Bioconductor[64] packages MSnbase[65] and pRoloc[25] for the R statistical programming language[66] were used for handling of the quantitative proteomics data and the protein-localization prediction. We employed the use of the full pRoloc pipeline in which proteins are assigned to a subcellular localization using a multi-step analysis framework.

Using the pRoloc software[25], an initial application of novelty detection[26] was conducted to identify and confirm the presence of organelle clusters in an unbiased data-specific manner. The novelty detection analysis was run as described in Breckels *et al.*[26] using the classic Gaussian ellipsoidal mixture models for multivariate data, a stringent 200 iteration run ($N = 200$), and outlier detection testing at the 5% level ($P = 0.05$). The minimum number of proteins per new phenotype cluster was set to 20 proteins (GS = 20), to allow detection of small organelles and complexes. A set of well-known residents from three distinct organelle structure; the mitochondria, plasma membrane and ER, and from three well-known and abundant protein complexes; the proteasome and two ribosomal subunits, 40S and 60S, were used as initial input markers for the discovery analysis (Supplementary Table 4). These initial markers were manually curated using information from the UniProt database[67], the Gene Ontology[68] and the literature. From the nature of the experimental design it was known that nuclear structures existed within the data, however, nuclear markers were left unlabelled in the discovery analysis to allow an unbiased detection of any sub-nuclear clusters. Also, markers that cover the lysosome, peroxisome and endosome were also left unlabelled to obtain an unbiased data-specific confirmation of their presence. Supplementary Fig. 15 shows the results of the discovery analysis and Supplementary Tables 5 and 6 show the number of clusters identified and the final marker set to be used in protein classification. This final set contained 13 different subcellular structures; the mitochondria, ER, plasma membrane, lysosome, peroxisome, endosome, actin cytoskeleton, extracellular matrix, nucleus (non-chromatin), chromatin, proteasome, 40S and 60S ribosomal subunits and the cytosol which were defined through a careful manual search of the literature and though validation of clusters from phenotype discovery analysis. These curated marker lists are available in the pRoloc software.

A SVM classifier, with a radial basis function kernel, using class specific weights was used for classification of unassigned proteins to one of the 14 known classes. The weights used in classification were set to be inversely proportional to the subcellular class frequencies to account for class imbalance. Algorithmic performance of the SVM on the data set was estimated using stratified fivefold cross-validation (creating five test/train partitions), which features an additional cross-validation on each training partition to optimize free parameters, sigma and cost, via a grid search (as described in ref. 23). This process was repeated 100 times and the best cost and sigma parameters were chosen based on the best F1 score; the harmonic mean of precision and recall. The best sigma was 0.01, which controls the bandwidth of the Gaussian, and the best cost was 16, which controls the balance between adherence to the training data and predictive performance on future unknown examples. As different organelles reflect different SVM score distributions (Supplementary Fig. 16), scoring thresholds were calculated per subcellular niche and were set based on concordance with gene ontology annotation to attain a 5% FDR. Unassigned proteins were then classified to 1 of the 13 compartments according to the SVM prediction if greater than the calculated class threshold. See also Supplementary Note 1.

Visualization and annotation of spatial proteomics data. To enable straightforward access to the data, we have developed the pRolocGUI application (http://bioconductor.org/packages/devel/bioc/html/pRolocGUI.html), which provides an interactive visualization interface for spatial proteomics data. It employs modern JavaScript technology that directly interacts with the R data using the shiny Web application framework for R (http://shiny.rstudio.com/). The interface enables users to visualize annotated spatial proteomics data with PCA and protein profile plots, search for proteins of interest and overlay protein complexes and functional networks onto the subcellular map. The application can be used to specifically explore our mouse pluripotent stem cell data online (https://lgatto.shinyapps.io/christoforou2015/), or can be installed locally to visualize data from any spatial proteomics experimental designs with or without any analysis with the pRoloc pipeline. Documentation and a series of online tutorial videos for pRolocGUI can be found at http://ComputationalProteomicsUnit.github.io/pRolocGUI/.

References

1. Bradley, A., Evans, M., Kaufman, M. H. & Robertson, E. Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. *Nature* **309,** 255–256 (1984).
2. Smith, A. Embryo-derived stem cells: of mice and men. *Annu. Rev. Cell Dev. Biol.* **17,** 435–462 (2001).
3. Chambers, I. & Tomlinson, S. R. The transcriptional foundation of pluripotency. *Development* **136,** 2311–2322 (2009).
4. Festuccia, N., Osorno, R., Wilson, V. & Chambers, I. The role of pluripotency gene regulatory network components in mediating transitions between pluripotent cell states. *Curr. Opin. Genet. Dev.* **23,** 504–511 (2013).
5. Sampath, P. *et al.* A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation. *Cell Stem Cell* **2,** 448–460 (2008).
6. Buckley, S. M. *et al.* Regulation of pluripotency and cellular reprogramming by the ubiquitin proteasome system. *Cell stem cell* **11,** 783–798 (2012).
7. Muñoz Descalzo, S. *et al.* A competitive protein interaction network buffers Oct4-mediated differentiation to promote pluripotency in embryonic stem cells. *Mol. Syst. Biol.* **9,** 694–694 (2013).

8. Shyh-Chang, N. & Daley, G. Q. Lin28: primal regulator of growth and metabolism in stem cells. *Cell stem cell* **12,** 395–406 (2013).
9. Uhlen, M. *et al.* Towards a knowledge-based human protein atlas. *Nat. Biotechnol.* **28,** 1248–1250 (2010).
10. Stadler, C. *et al.* Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. *Nat Methods* **10,** 315–323 (2013).
11. Christoforou, A., Mulvey, C. M., Breckels, L. M., Gatto, L. & Lilley, K. S. in *Quantitative Proteomics New Developments in Mass Spectrometry* (eds Eyers, Claire E. & Gaskell, Simon) Ch. 9 185–210 (Royal Society of Chemistry, 2014).
12. Dunkley, T. P. J. *et al.* Mapping the arabidopsis organelle proteome. *Proc. Natl Acad. Sci. USA* **103,** 6518–6523 (2006).
13. Thompson, A. *et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75,** 1895–1904 (2003).
14. de Duve, C. Tissue fractionation. *J. Cell Biol.* **50,** 20D–55D (1971).
15. McAlister, G. C. *et al.* Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal. Chem.* **84,** 7469–7478 (2012).
16. Christoforou, A. & Lilley, K. Isobaric tagging approaches in quantitative proteomics: the ups and downs. *Anal. Bioanal. Chem.* **404,** 1029–1037 (2012).
17. Ting, L., Rad, R., Gygi, S. G. & Haas, W. MS3 eliminates ratio distortion in isobaric labeling multiplexed quantitative proteomics. *Nat. Methods* **8,** 937–940 (2011).
18. Wenger, C. D. *et al.* Gas-phase purification enables accurate, large-scale, multiplexed proteome quantification. *Nat. Methods* **8,** 933–935 (2011).
19. Dayon, L., Sonderegger, B. & Kussmann, M. Combination of gas-phase fractionation and MS3 acquisition modes for relative protein quantification with isobaric tagging. *J. Proteome Res.* **11,** 5081–5089 (2012).
20. Vachet, R. W. & McElvany, S. W. Application of external customized waveforms to a commercial quadrupole ion trap. *J. Am. Soc. Mass Spectrom.* **10,** 355–359 (1999).
21. McAlister, G. C. *et al.* MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* **86,** 7150–7158 (2014).
22. Ow, S. Y. *et al.* iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". *J. Proteome Res.* **8,** 5347–5355 (2009).
23. Trotter, M. W. B., Sadowski, P. G., Dunkley, T. P. J., Groen, A. J. & Lilley, K. S. Improved sub-cellular resolution via simultaneous analysis of organelle proteomics data across varied experimental conditions. *Proteomics* **10,** 4213–4219 (2010).
24. Groen, A. J. *et al.* Identification of trans-Golgi network proteins in arabidopsis thaliana root tissue. *J. Proteome Res.* **13,** 763–776 (2013).
25. Gatto, L., Breckels, L. M., Wieczorek, S., Burger, T. & Lilley, K. S. Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics* **30,** 1322–1324 (2014).
26. Breckels, L. M. *et al.* The effect of organelle discovery upon sub-cellular protein localisation. *J. Proteomics* **88,** 129–140 (2013).
27. Bausch-Fluck, D. *et al.* A Mass spectrometric-derived cell surface protein atlas. *PLoS One* **10,** e0121314 (2015).
28. Faunes, F. *et al.* A membrane-associated β-catenin/Oct4 complex correlates with ground-state pluripotency in mouse embryonic stem cells. *Development* **140,** 1171–1183 (2013).
29. Krajewski, S. *et al.* Immunohistochemical analysis of Mcl-1 and Bcl-2 proteins in normal and neoplastic lymph nodes. *Am. J. Pathol.* **145,** 515–525 (1994).
30. Betschinger, J. *et al.* Exit from pluripotency is gated by intracellular redistribution of the bHLH transcription factor Tfe3. *Cell* **153,** 335–347 (2013).
31. Moremen, K. W. & Robbins, P. W. Isolation, characterization, and expression of cDNAs encoding murine alpha-mannosidase II, a Golgi enzyme that controls conversion of high mannose to complex N-glycans. *J. Cell Biol.* **115,** 1521–1534 (1991).
32. Lowe, S. L., Peter, F., Subramaniam, V. N., Wong, S. H. & Hong, W. A SNARE involved in protein transport through the Golgi apparatus. *Nature* **389,** 881–884 (1997).
33. Kim, H. W. *et al.* Human macrophage ATP7A is localized in the trans-Golgi apparatus, controls intracellular copper levels, and mediates macrophage responses to dermal wounds. *Inflammation* **35,** 167–175 (2012).
34. Pevsner, J., Hsu, S.-C., Hyde, P. S. & Scheller, R. H. Mammalian homologues of yeast vacuolar protein sorting (vps) genes implicated in Golgi-to-lysosome trafficking. *Gene* **183,** 7–14 (1996).
35. Stehbens, S. J. *et al.* CLASPs link focal-adhesion-associated microtubule capture to localized exocytosis and adhesion site turnover. *Nat. Cell Biol.* **16,** 558–570 (2014).
36. Ohta, E. *et al.* Identification and characterization of GCP16, a novel acylated golgi protein that interacts with GCP170. *J. Biol. Chem.* **278,** 51957–51967 (2003).
37. Kodani, A. & Sütterlin, C. The Golgi protein GM130 regulates centrosome morphology and function. *Mol. Biol. Cell.* **19,** 745–753 (2008).
38. Andersen, J. S. *et al.* Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426,** 570–574 (2003).
39. Nogales-Cadenas, R., Abascal, F., Díez-Pérez, J., Carazo, J. M. & Pascual-Montano, A. CentrosomeDB: a human centrosomal proteins database. *Nucleic Acids Res.* **37,** D175–D180 (2009).
40. Jakobsen, L. *et al.* Novel asymmetrically localizing components of human centrosomes identified by complementary proteomics methods. *EMBO J.* **30,** 1520–1535 (2011).
41. Sütterlin, C. & Colanzi, A. The Golgi and the centrosome: building a functional partnership. *J. Cell Biol.* **188,** 621–628 (2010).
42. Turner, D. A., Trott, J., Hayward, P., Rué, P. & Martinez Arias, A. An interplay between extracellular signalling and the dynamics of the exit from pluripotency drives cell fate decisions in mouse ES cells. *Biol. Open* **3,** 614–626 (2014).
43. Coronado, D. *et al.* A short G1 phase is an intrinsic determinant of naïve embryonic stem cell pluripotency. *Stem Cell Res.* **10,** 118–131 (2013).
44. Tan, D. J. L. *et al.* Mapping organelle proteins and protein complexes in drosophila melanogaster. *J. Proteome Res.* **8,** 2667–2678 (2009).
45. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42,** D199–D205 (2014).
46. Milacic, M. *et al.* Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers* **4,** 1180–1211 (2012).
47. Croft, D. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **42,** D472–D477 (2014).
48. Gegonne, A. *et al.* TFIID component TAF7 functionally interacts with both TFIIH and P-TEFb. *Proc. Natl Acad. Sci. USA* **105,** 5367–5372 (2008).
49. Das, A. *et al.* RalA promotes a direct exocyst–Par6 interaction to regulate polarity in neuronal development. *J. Cell Sci.* **127,** 686–699 (2014).
50. Havugimana, P. C. *et al.* A census of human soluble protein complexes. *Cell* **150,** 1068–1081 (2012).
51. van den Berg, D. L. C. *et al.* An Oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell* **6,** 369–381 (2010).
52. Gao, Z. *et al.* Determination of protein interactome of transcription factor sox2 in embryonic stem cells engineered for inducible expression of four reprogramming factors. *J. Biol. Chem.* **287,** 11384–11397 (2012).
53. Gagliardi, A. *et al.* A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *EMBO J.* **32,** 2231–2247 (2013).
54. Liang, J. *et al.* Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells. *Nat. Cell Biol.* **10,** 731–739 (2008).
55. Ahmad, Y., Boisvert, F.-M., Lundberg, E., Uhlen, M. & Lamond, A. I. Systematic analysis of protein pools, isoforms, and modifications affecting turnover and subcellular localization. *Mol. Cell. Proteomics* **11** 1–11 doi:10.1074/mcp.M111.013680 (2012).
56. Horton, P. *et al.* WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* **35,** W585–W587 (2007).
57. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41,** D808–D815 (2013).
58. Mohan, K. N., Ding, F. & Chaillet, J. R. Distinct roles of DMAP1 in mouse development. *Mol. Cell. Biol.* **31,** 1861–1869 (2011).
59. Grohmann, M. *et al.* Restricted mobility of Dnmt1 in preimplantation embryos: implications for epigenetic reprogramming. *BMC Dev. Biol.* **5,** 18–18 (2005).
60. Kalmar, T. *et al.* Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol.* **7,** e1000149 (2009).
61. Wysocka, J., Reilly, P. T. & Herr, W. Loss of HCF-1-chromatin association precedes temperature-induced growth arrest of tsBN67 cells. *Mol. Cell. Biol.* **21,** 3820–3829 (2001).
62. Wühr, M. *et al.* Accurate multiplexed proteomics at the MS2 level using the complement reporter ion cluster. *Anal. Chem.* **84,** 9214–9221 (2012).
63. Brosch, M., Yu, L., Hubbard, T. & Choudhary, J. Accurate and sensitive peptide identification with mascot percolator. *J. Proteome Res.* **8,** 3176–3181 (2009).
64. Gentleman, R. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).
65. Gatto, L. & Lilley, K. S. MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics* **28,** 288–289 (2012).
66. R Core Team, R. *R: A language and environment for statistical computing. R Foundation for Statistical Computing,* http://www.R-project.org/ (2013).
67. The UniProt, C. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **40,** D71–D75 (2012).
68. The Reference Genome Group of the Gene Ontology, C. The gene ontology's reference genome project: a unified framework for functional annotation across species. *PLoS Comput. Biol.* **5,** e1000431 (2009).
69. Vizcaino, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32,** 223–226 (2014).

Acknowledgements

We thank Andreas Hühmer, Philip Remes, Jesse Canterbury and Graeme McAlister of Thermo Fisher Scientific, San Jose, California, USA, for their advice regarding operation of the Orbitrap Fusion. We also thank Mike Deery for assistance with checking sample integrity on the mass spectrometers in the Cambridge Centre for Proteomics on equipment purchased via a Wellcome Trust grant (099135/Z/12/Z), and Brian Hendrich of the Wellcome Trust-MRC Stem Cell Institute in Cambridge and Sean Munro of the MRC Laboratory of Molecular Biology in Cambridge for insightful comments about the data. A.C. was supported by BBSRC grant (BB/D526088/1). C.M.M. and L.G. were supported by European Union 7th Framework Program (PRIMEXS project, grant agreement number 262067), L.M.B was supported by a BBSRC Tools and Resources Development Fund (Award BB/K00137X/1), and P.C.H. was supported by an ERC Advanced Investigator grant to A.M.A. A.G. was funded through the Alexander S. Onassis Public Benefit Foundation, the Foundation for Education and European Culture (IPEP) and the Embiricos Trust Scholarship of Jesus College Cambridge. T.H. was supported by Commonwealth Split Site PhD Scholarship. T.N. was supported by an ERASMUS Placement scholarship.

Author contributions

A.C., C.M.M., A.M.A. and K.S.L. designed and planned the experiments. A.C., C.M.M., A.G., T.H. and P.C.H. performed the cell culture and sample preparation. A.C., C.M.M. and R.V. performed the mass spectrometry. A.C., L.M.B. and L.G. performed the multivariate data analysis. L.M.B., T.N. and L.G. created tools and tutorials for data visualization. All authors contributed to the technical and biological interpretation of data, and to the writing of the manuscript.

Additional information

Accession codes: The mass spectrometry data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org)[69] via the PRIDE partner repository with the data set identifier PXD001279.

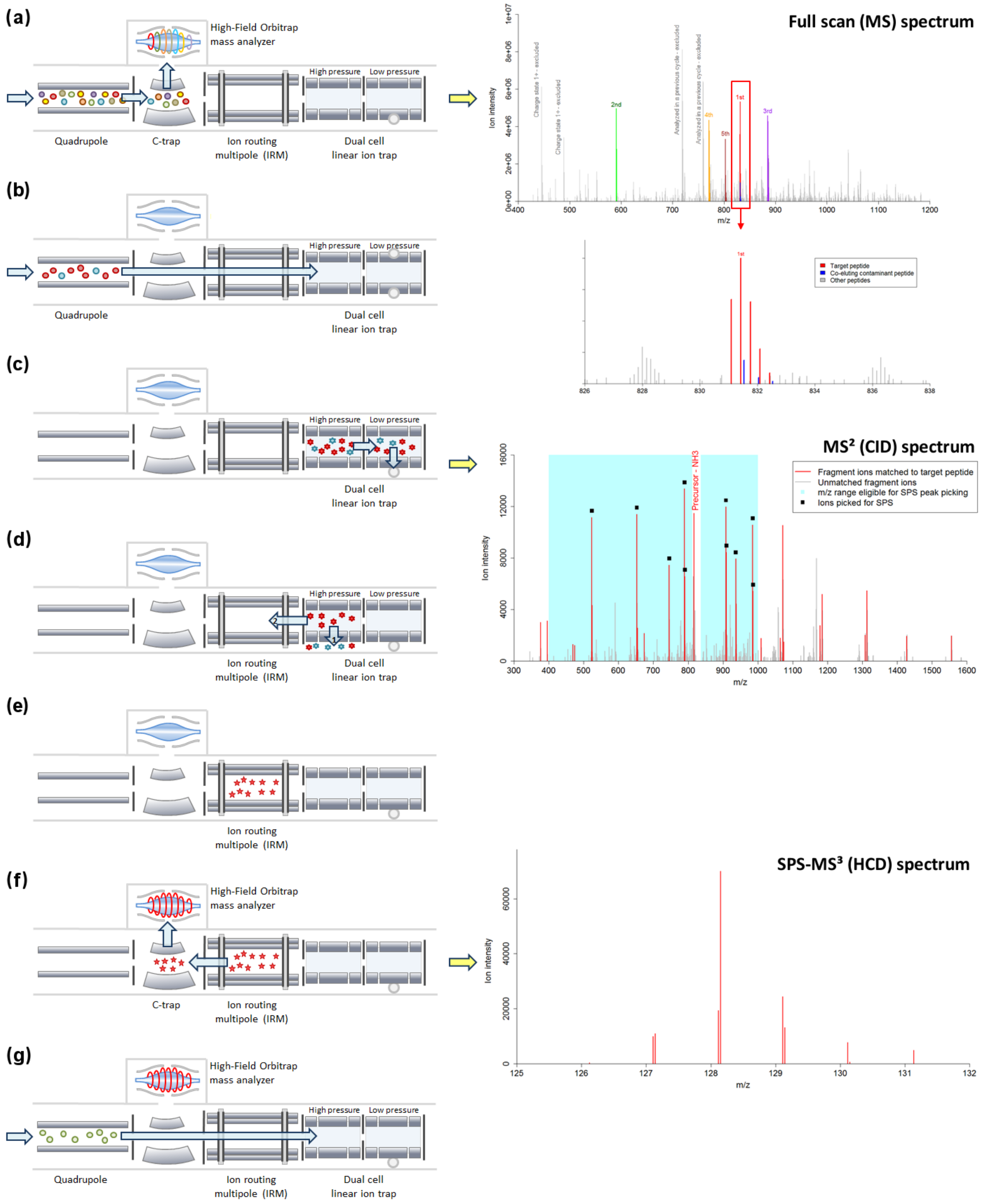
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

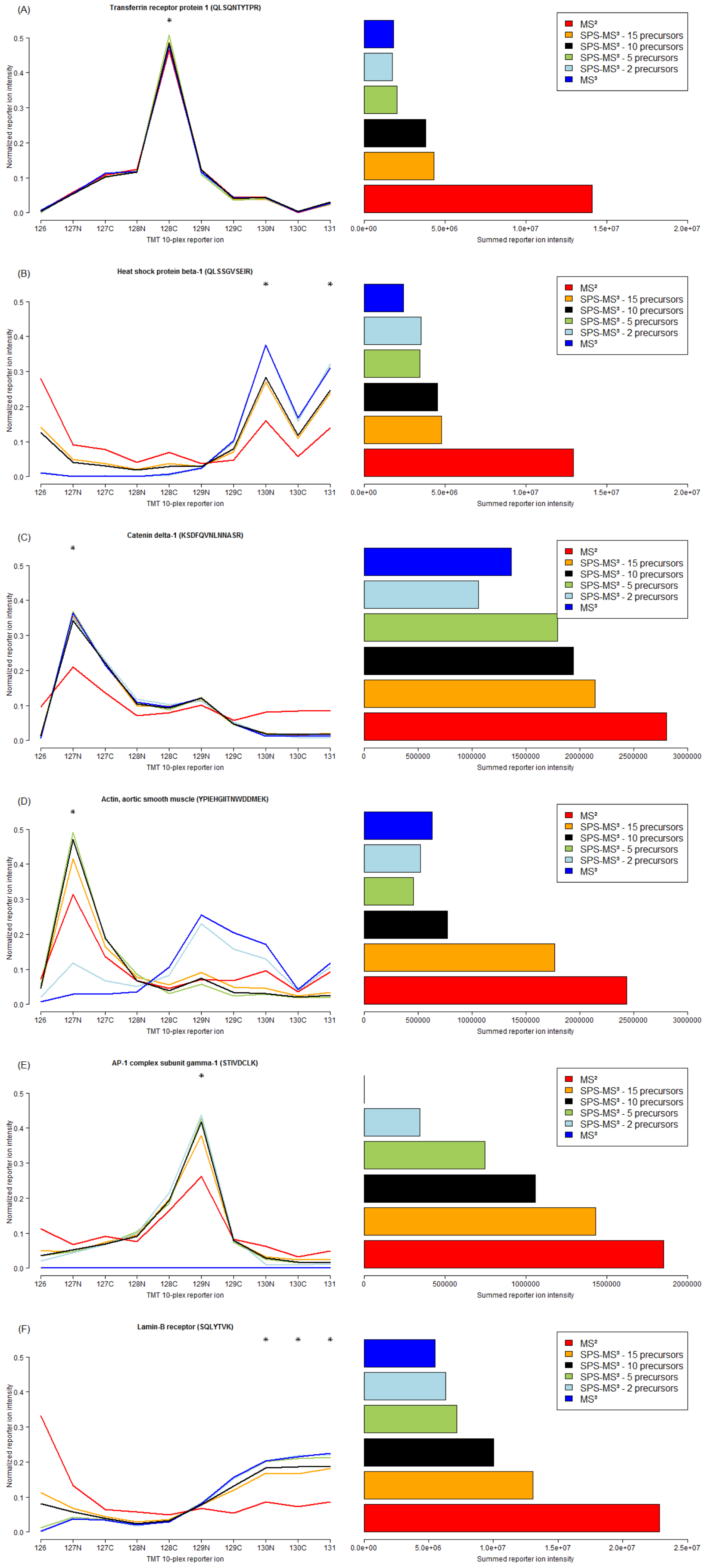
Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Christoforou, A. *et al.* A draft map of the mouse pluripotent stem cell spatial proteome. *Nat. Commun.* 7:8992 doi: 10.1038/ncomms9992 (2016).

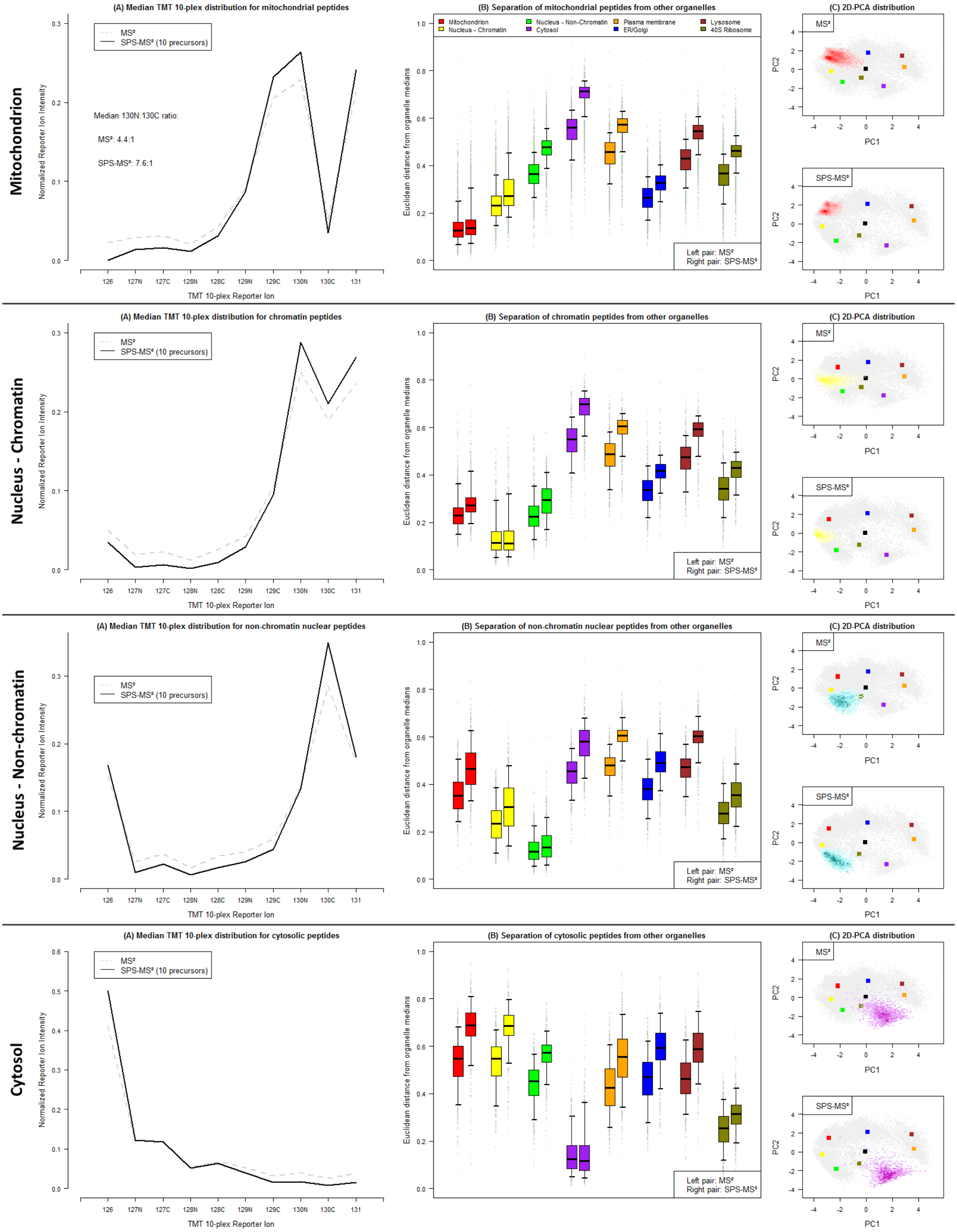
Supplementary Information



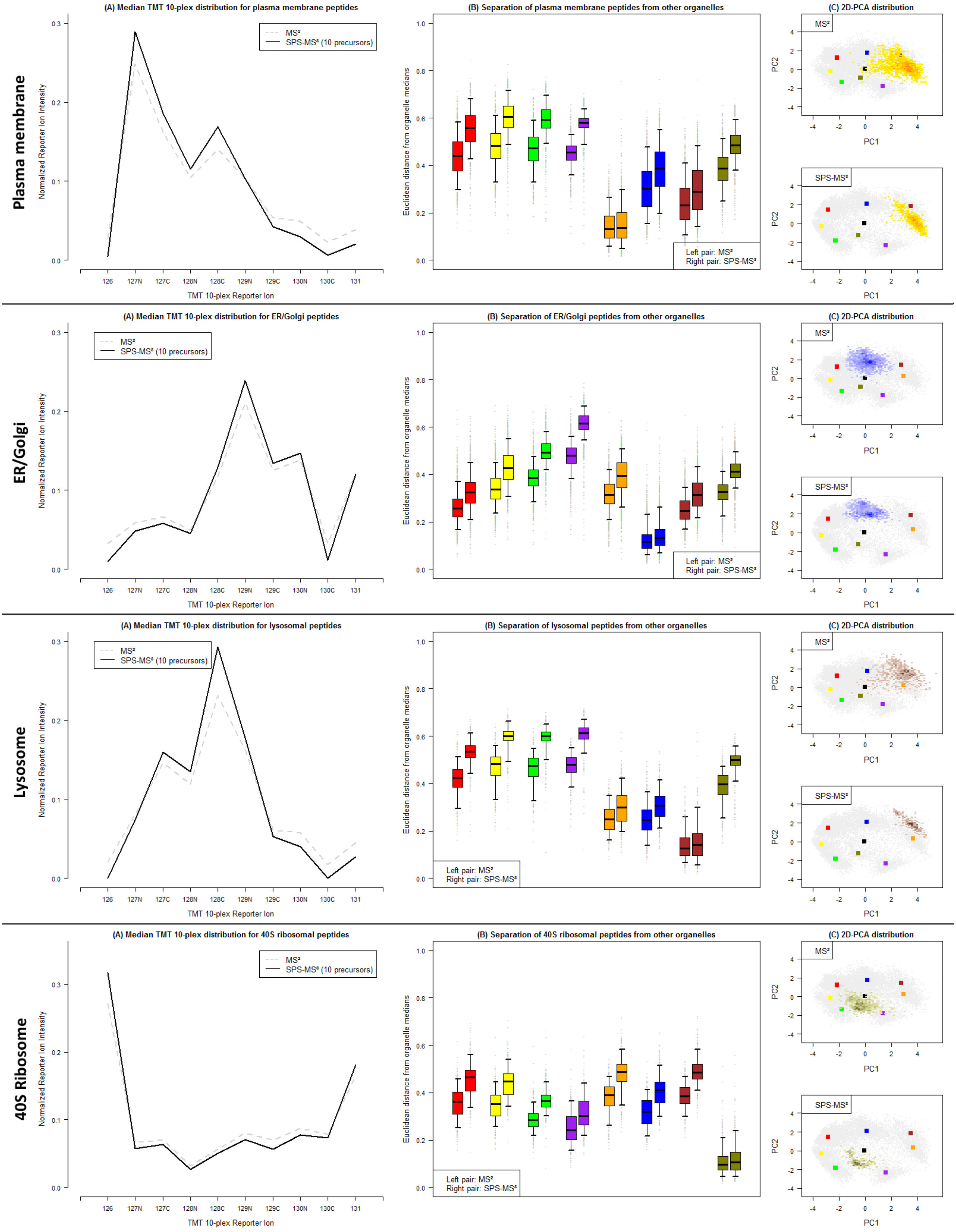
Supplementary Figure 1 | Overview of steps performed in SPS-MS^3 acquisition. (a) Peptides eluting from the LC system undergo electrospray ionization. Cations are then transmitted through a quadrupole and accumulated in a C-trap, before being passed into the Orbitrap mass analyzer to generate a full scan (MS) spectrum. The most intense ions detected in the full scan (5 in this example) are selected for SPS-MS^3 analysis. (b) The first precursor selected in the full scan (red) is isolated and transmitted to the linear ion trap via the ion routing multipole (IRM). Due to the high sample complexity, other peptides with similar *m/z* and LC retention properties are co-selected (blue). (c) Peptides are fragmented by collision induced dissociation (CID) in the high pressure cell of the ion trap. Fragments are then transferred to the low pressure cell and scanned out onto two electron multipliers to generate an MS^2 CID spectrum. The most intense ions in this spectrum that meet specified criteria (as described in the Methods section), which are typically specific to the target peptide, are determined on-the-fly and picked for SPS. The MS^2 spectrum is also used post-acquisition for peptide identification. (d) Steps b and c are repeated for the same precursor. Rather than transferring peptide fragments to the low pressure cell after fragmentation, an SPS isolation waveform is applied to eject all ions from the linear ion trap except those chosen for SPS (1). The purified SPS ions are then transferred to the IRM (2). (e) The SPS ions undergo a second round of fragmentation by higher-energy collisional dissociation (HCD), liberating TMT reporter ions from the labeled peptide fragments. (f) These fragments are transferred via the C-trap into the Orbitrap to generate an SPS-MS^3 HCD spectrum, which is used for relative quantification of TMT 10-plex reporter ions in the *m/z* range of 126-131. (g) In order to utilize analysis time efficiently, operation of the Orbitrap and linear ion trap is parallelized. For example, while the Orbitrap is acquiring the SPS-MS^3 for the first precursor (red), the second precursor (green) is already being isolated by the quadrupole and fragmented in the linear ion trap to generate its MS^2 spectrum.



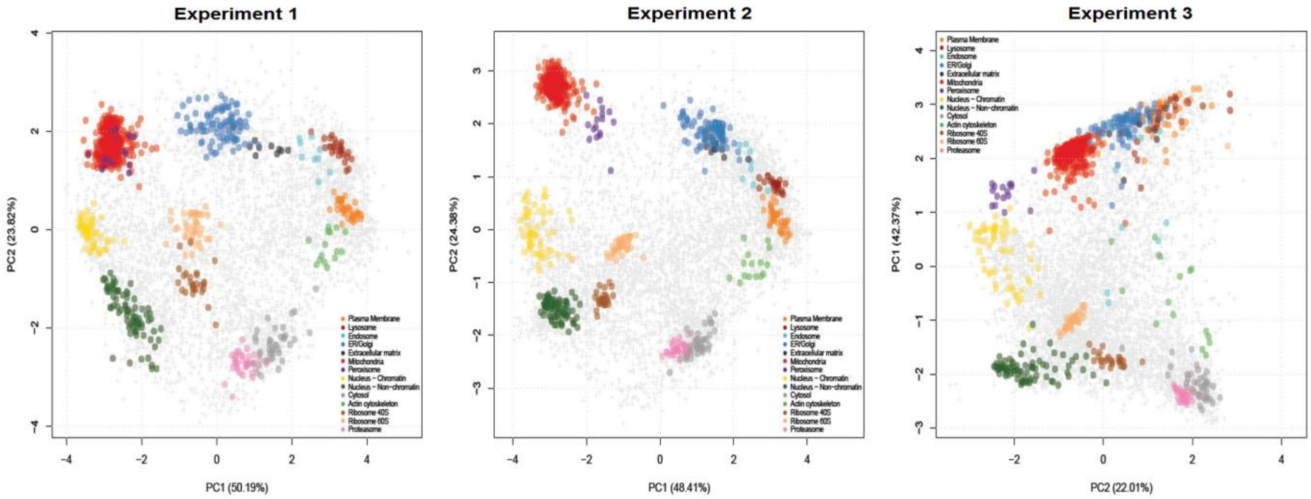
Supplementary Figure 2 | Examples of the differences observed with conventional MS^2, conventional MS^3, and $SPS-MS^3$ acquisition for Experiment 1. Plots on the left side demonstrate differences in quantitative accuracy, with the expected peak fraction(s) for a particular peptide represented by asterisks. Plots on the right show the effects of acquisition mode on summed reporter ion signal. (A) No change in quantitative performance observed. (B) $SPS-MS^3$ provides quantitative accuracy between that of MS^2 and MS^3, with fewer precursors producing more accurate data. (C) $SPS-MS^3$ and conventional MS^3 are both more accurate than MS^2, and $SPS-MS^3$ improves reporter ion signal relative to MS^3. (D) MS^3 and $SPS-MS^3$ with few precursors yield inaccurate quantification due to the contaminant fragment ion(s) being selected for re-fragmentation. $SPS-MS^3$ with more precursors compensates for this. (E) Conventional MS^3 is unquantifiable due to poor signal. $SPS-MS^3$ restores signal whilst preserving gains in quantitative performance over MS^2. (F) Distorted MS^2 reporter ion distribution results in misclassification of protein localization. The expected distribution is restored by $SPS-MS^3$.



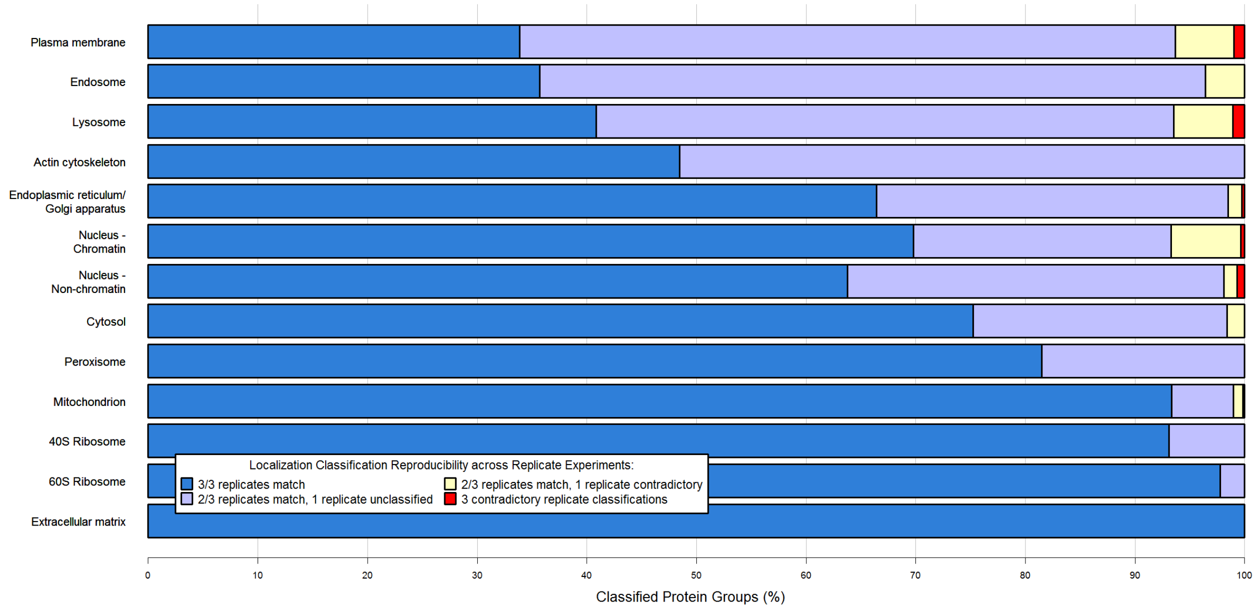
Supplementary Figure 3 | Enhanced organelle resolution with SPS-MS^3 relative to conventional MS^2 acquisition. Additional examples of the effect of SPS-MS^3 on organellar resolution to those presented in Figure 2.



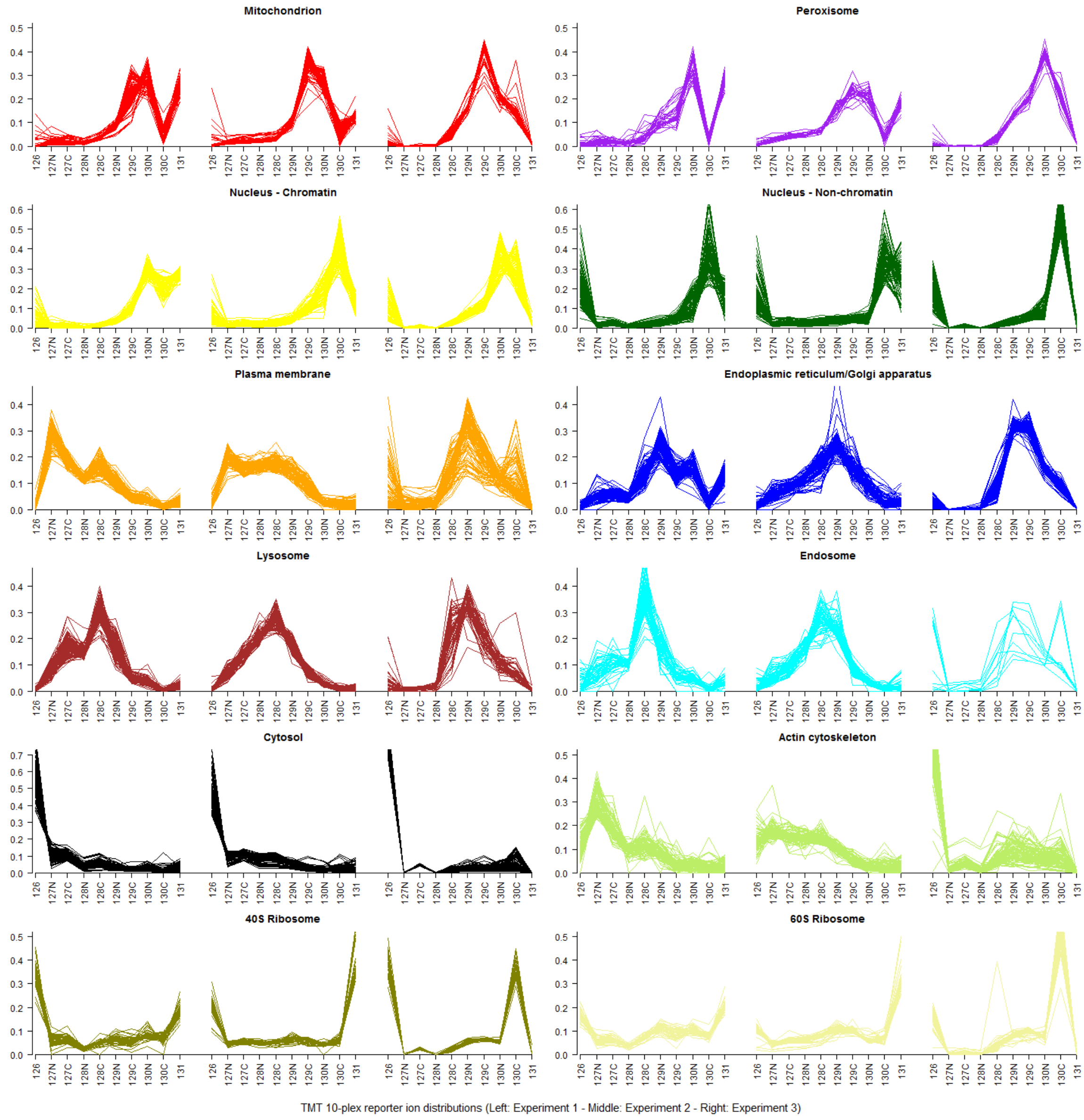
Supplementary Figure 3 (continued).



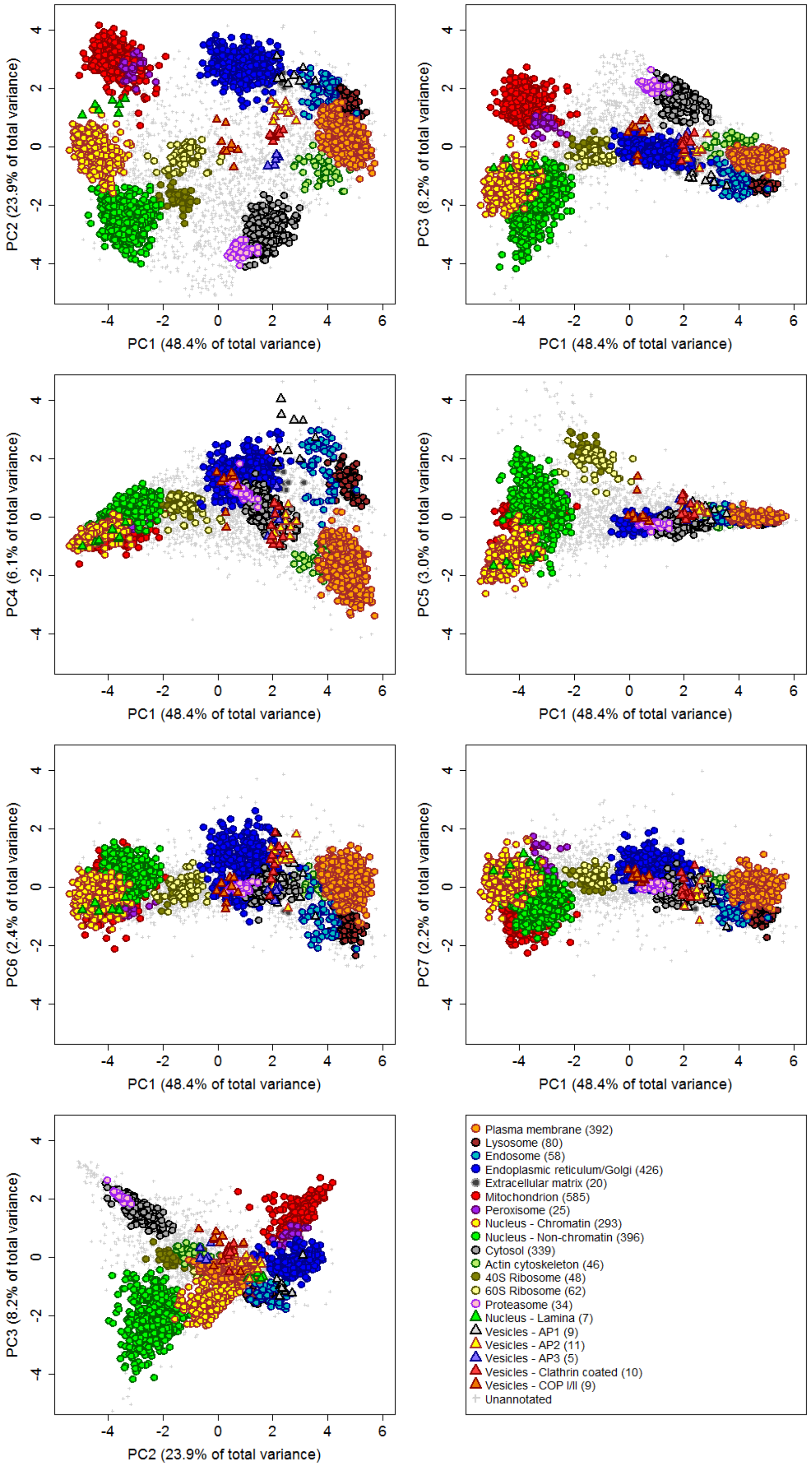
Supplementary Figure 4 | Principal component analysis plots for the three fractionation experiments. Organellar marker proteins are colored points, while other proteins appear as gray points.



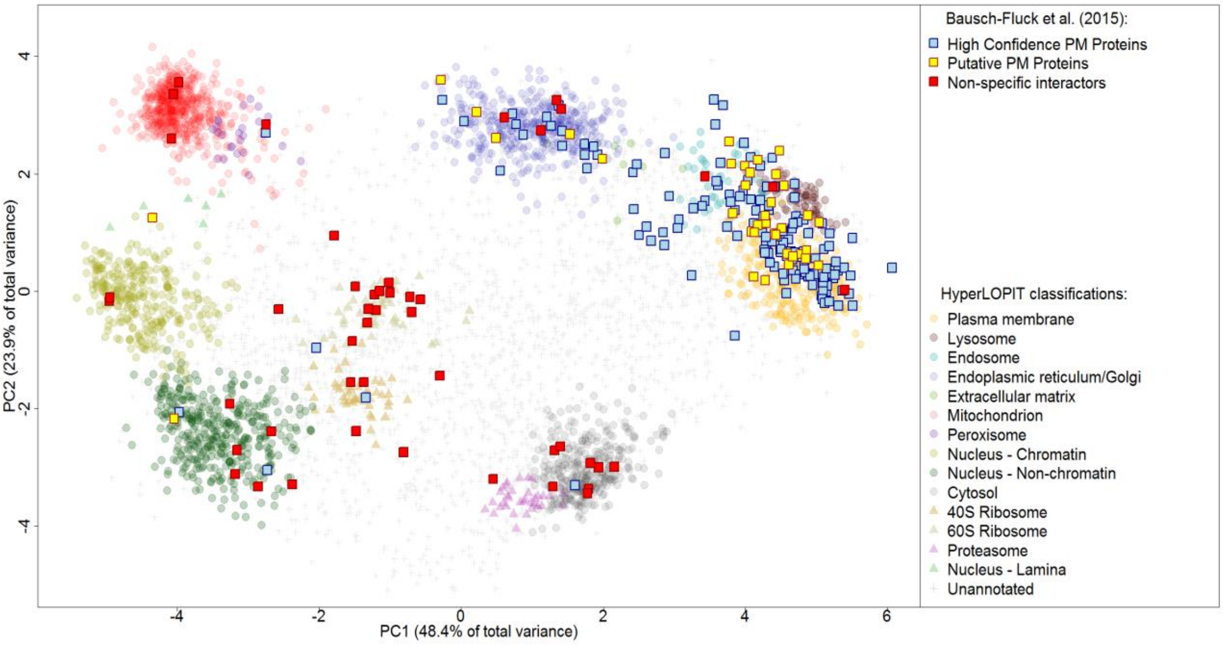
Supplementary Figure 5 | Concordance in protein localization assignments across three experiments. SVM classification was performed on each of the three experiments individually. 2,841 proteins were both identified in all three experiments, and assigned to an organelle class in at least 2 of 3 experiments. Over 90% of these proteins were either assigned to the same localization class in all 3 cases, or in 2 of 3 cases with the remaining experiment not classifying the protein to any of the 13 subcellular classes. Contradictory localization assignments occur at suitably low frequency for all organelle classes.



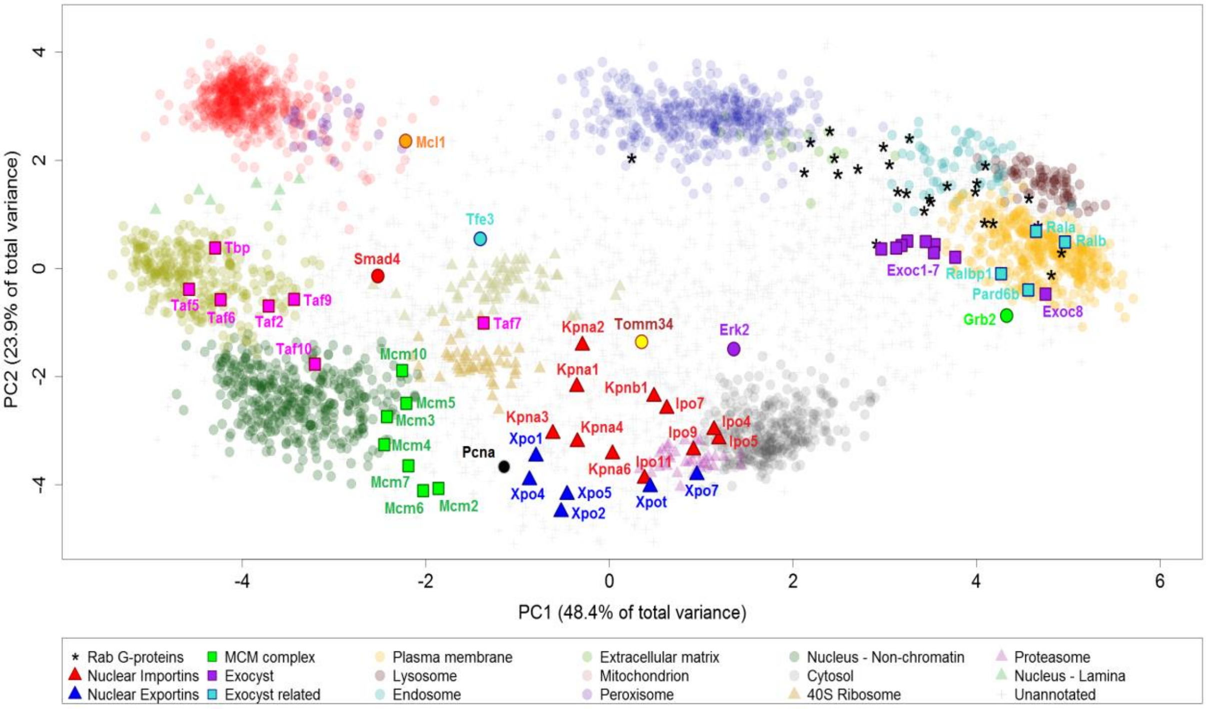
Supplementary Figure 6 | TMT 10-plex reporter ion distributions for two biological replicates of E14TG2a hyperLOPIT data. Proteins with the same steady-state subcellular localization co-distribute across the fractionation scheme and therefore show correlated and characteristic multivariate profiles.



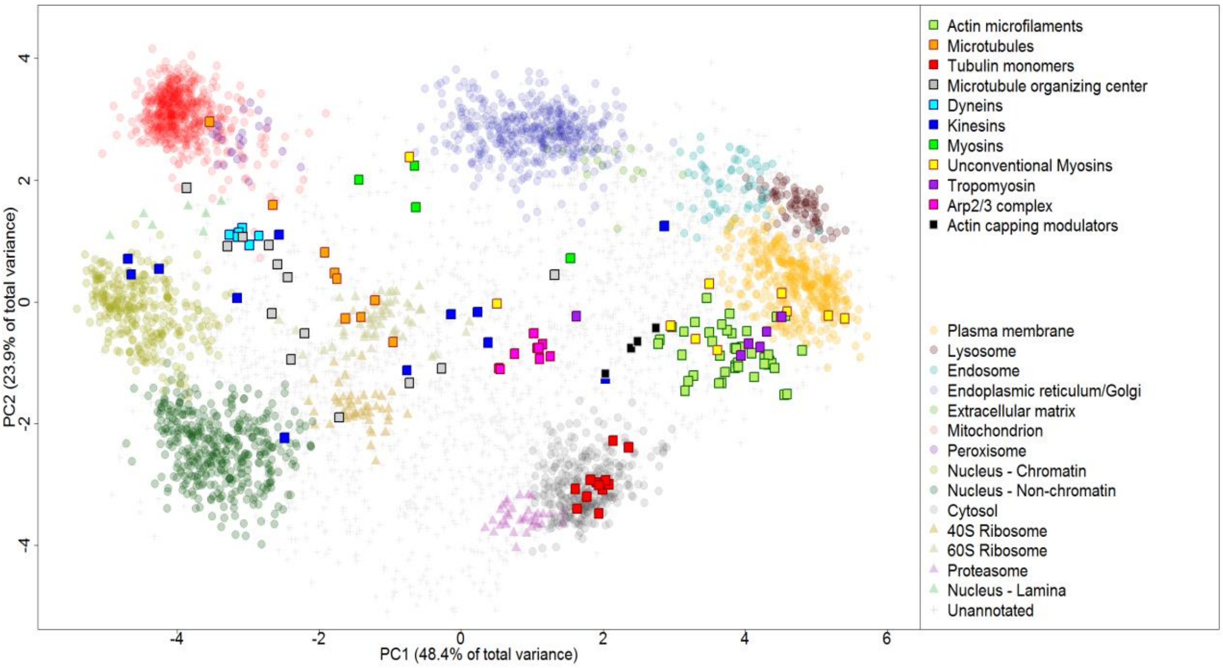
Supplementary Figure 7 | Lower principal components for hyperLOPIT data presented in Figure 3. Resolution of some compartments is more evident in the lower components, for example separation of plasma membrane and lysosome in PC3, and mitochondrion and peroxisome in PC7.



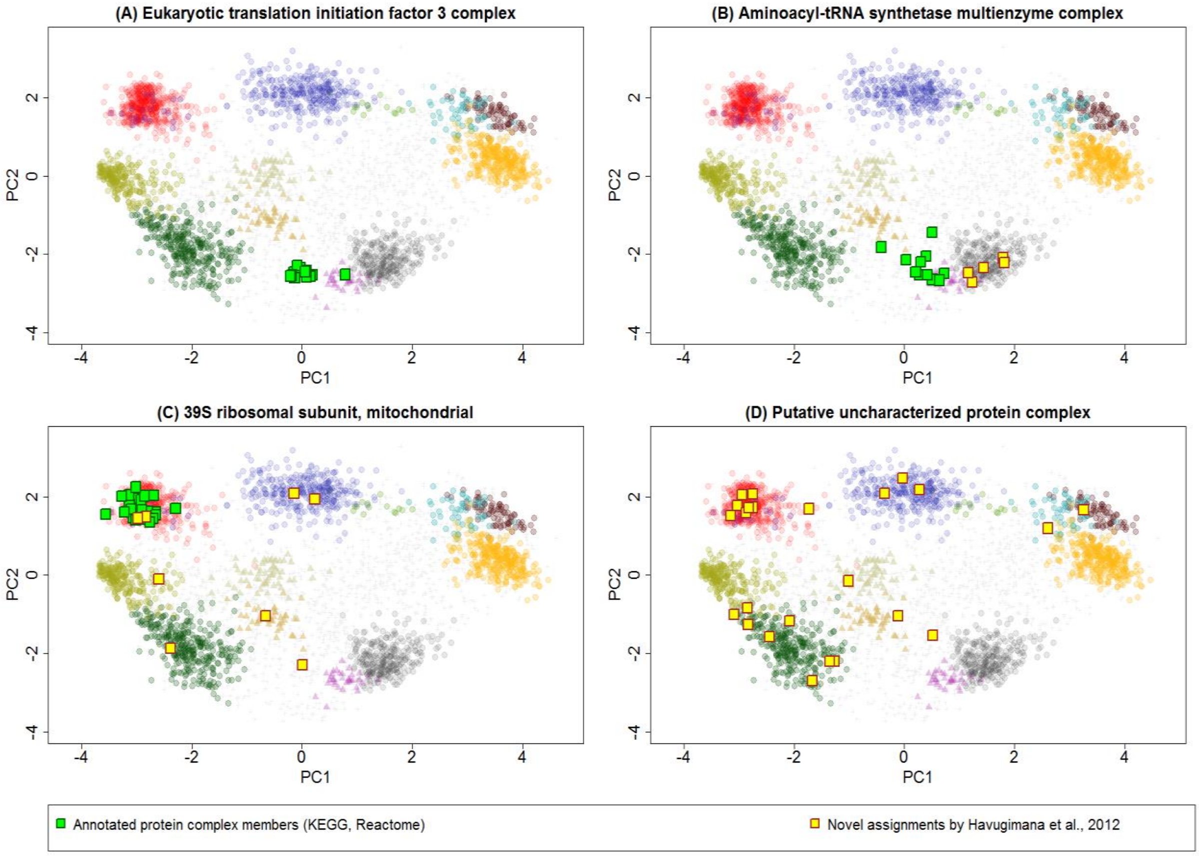
Supplementary Figure 8 | E14Tg2a cell surface capture data from Bausch-Fluck *et al.* [1] overlayed onto the hyperLOPIT dataset. HyperLOPIT verifies localization of high confidence cell surface proteins assigned by Bausch-Fluck and co-workers, and provides experimental evidence to support putative cell surface proteins. Most such proteins observed in the plasma membrane or trans-Golgi network in the hyperLOPIT data. Proteins described as non-specific interactors by Bausch-Fluck and co-workers display hyperLOPIT distributions that are inconsistent with cell surface proteins, such as mitochondrial, nuclear, ribosomal and cytosolic localization.



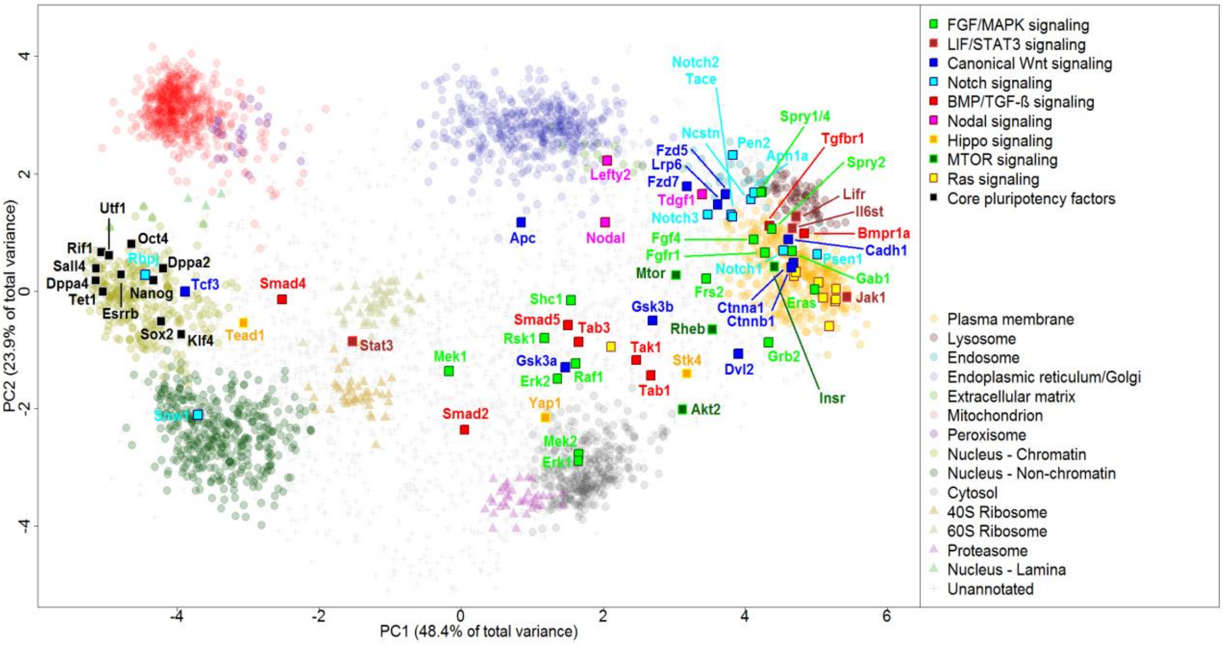
Supplementary Figure 9 | Examples of proteins displaying mixed localization. Proteins with mixed localization do not co-distribute with classifiable organelle phenotypes (muted colors), and therefore display characteristic distribution patterns. The nuclear import/export machinery demonstrates an intermediate position between the cytosol and the nucleus, while the Rab G-proteins are distributed throughout the secretory pathway. Similarly, the MCM (minichromosome maintenance) complex has a distinct location between the nucleus and cytoplasm, in accordance with its role in DNA replication initiation [2]. Tfe3 is a transcription factor whose nuclear/cytoplasmic ratio is indicative of differentiation status [3], while nucleocytoplasmic re-localization of Pcna has previously been demonstrated to modulate differentiation of neutrophils [4]. The Bcl-2 family member Mcl-1 displays an intermediate position between the mitochondria, endoplasmic reticulum and the nucleus [5]. Tom34 is a cytosolic co-chaperone involved in mitochondrial protein import [6]. Also shown are two examples of complexes where a single member of the complex has a distinct localization from the core group (TFIID complex and the exocyst complex). Taf7 is thought to dissociate from the TFIID complex following initiation of transcription, and also has a role in the assembly of several other transcription pre-initiation complexes, which may explain its separate steady-state location [7]. Exoc8 is localized away from the core exocyst complex, and co-distributes with its known binding partners Par6 and RalA [8].



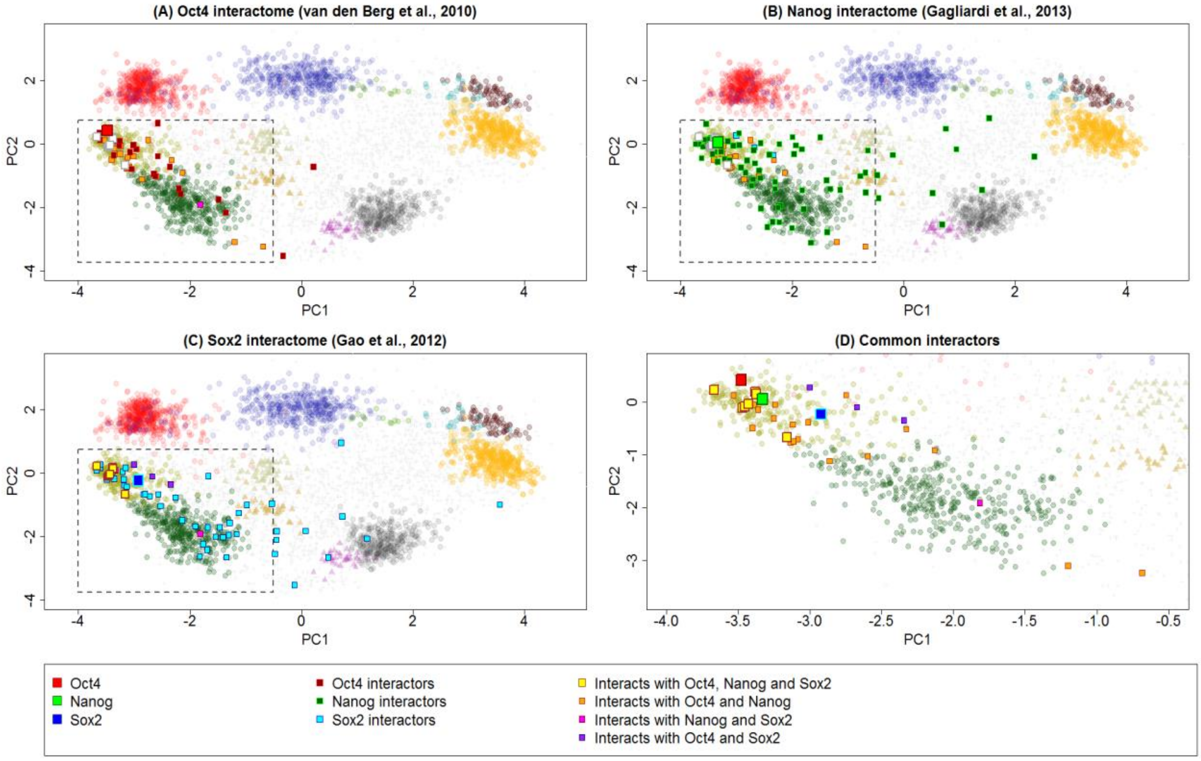
Supplementary Figure 10 | Fractionation of the cytoskeleton by hyperLOPIT. In addition to separating organelles and soluble proteins, cytoskeletal components also fractionate with characteristic distribution patterns.



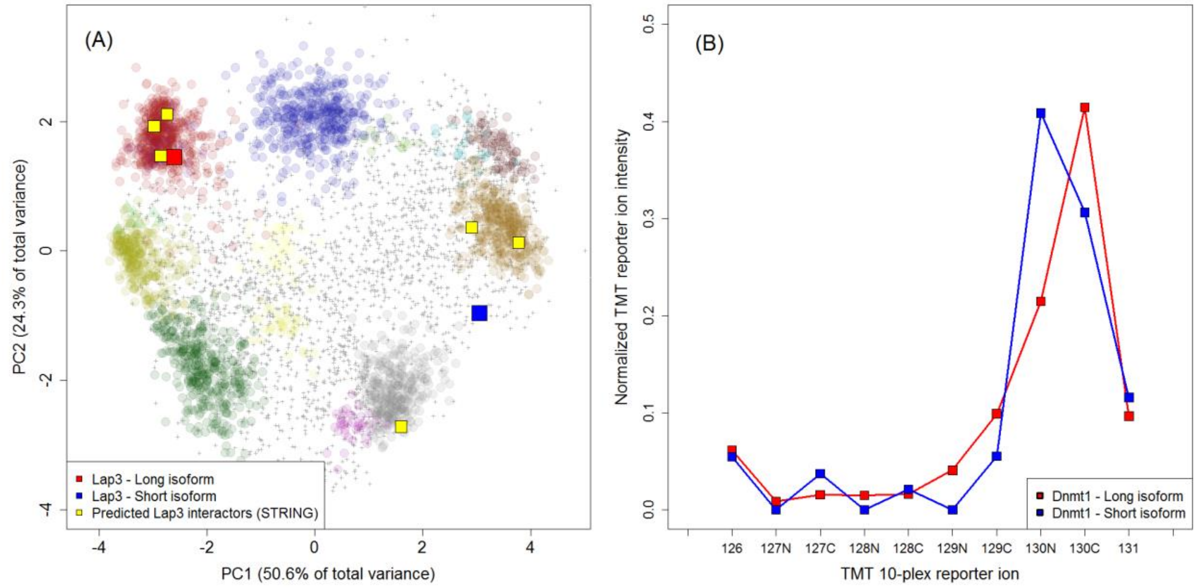
Supplementary Figure 11 | Mouse orthologues of protein complexes identified in a census of the human proteome. [9] (A) Components of highly curated and characterized complexes such as the eIF3 complex display closely correlated subcellular distribution profiles. (B) HyperLOPIT adds a spatial context to interactomics studies. The aminoacyl-tRNA synthetase complex is distributed between the cytosol and ribosome, consistent with its expected function. Additional assignments to the human orthologue of this complex by Havugimana and co-workers were localized to the cytosol, suggesting that their interaction with the complex is transient or unstable relative to the 'core' curated complex. (C) The spatial context can also be used to add additional confidence to novel assignments. Two of eight proteins novel assignments to the mitochondrial ribosome were found to localize to the mitochondrion and were therefore plausible interactors. The remaining six novel interactors were distributed across a range of other subcellular compartments, suggesting that these interactions are improbable. (D) Putative protein complexes can also be evaluated with this approach. Components of the putative complex shown here are distributed in many different subcellular compartments, suggesting that the probability of all components co-localizing to form a single complex is low. The putative complex is therefore likely a false positive in this case.



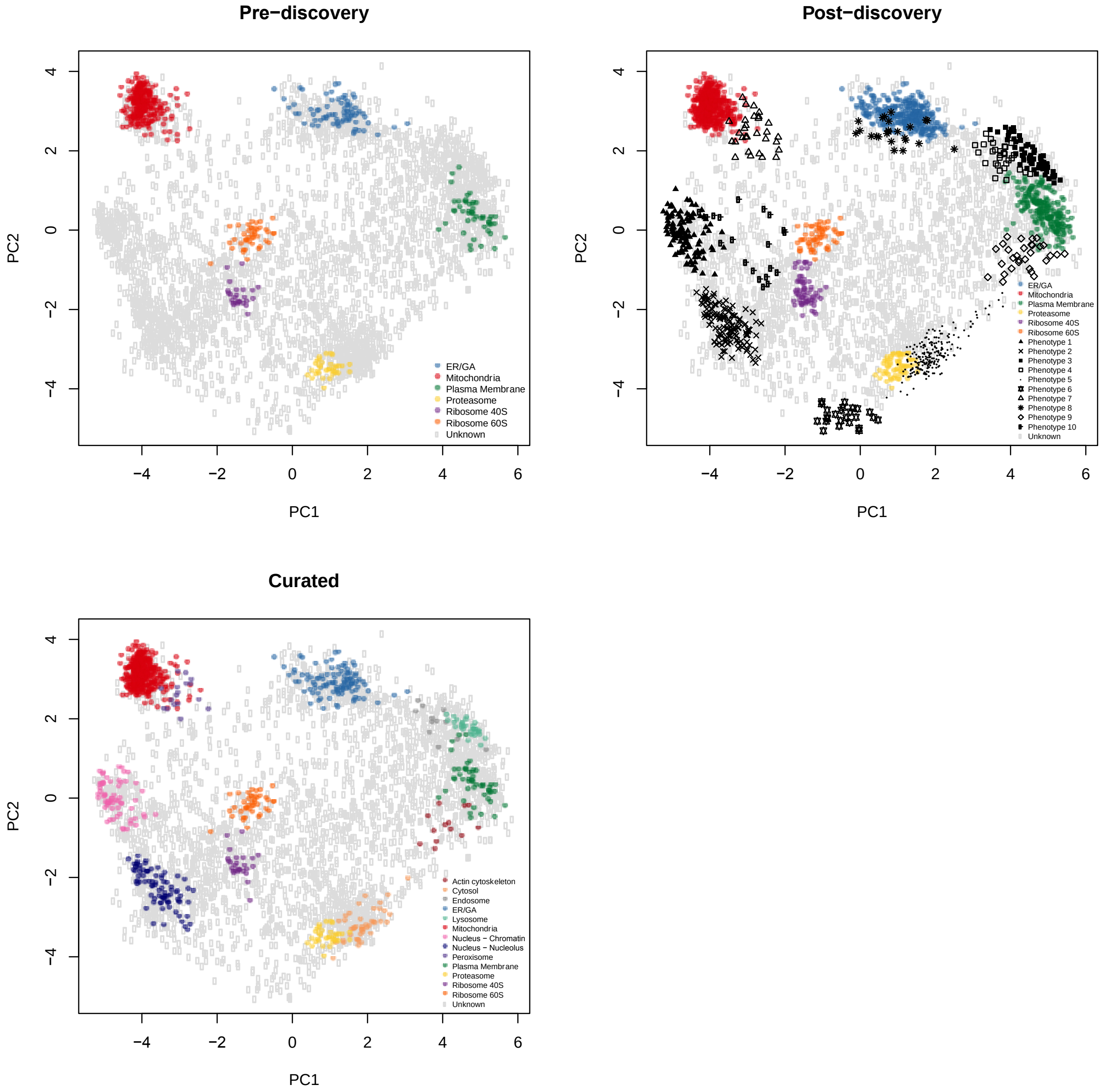
Supplementary Figure 12 | Subcellular localization of signaling cascade components. Pluripotency is maintained by a network of transcription factors which are influenced by several well-defined signaling pathways in combination with intrinsic and extrinsic factors. Six key signaling cascades involved in cell fate determination are highlighted, including cell surface receptors and ligands, intracellular adapters and transducers, and nuclear effectors.



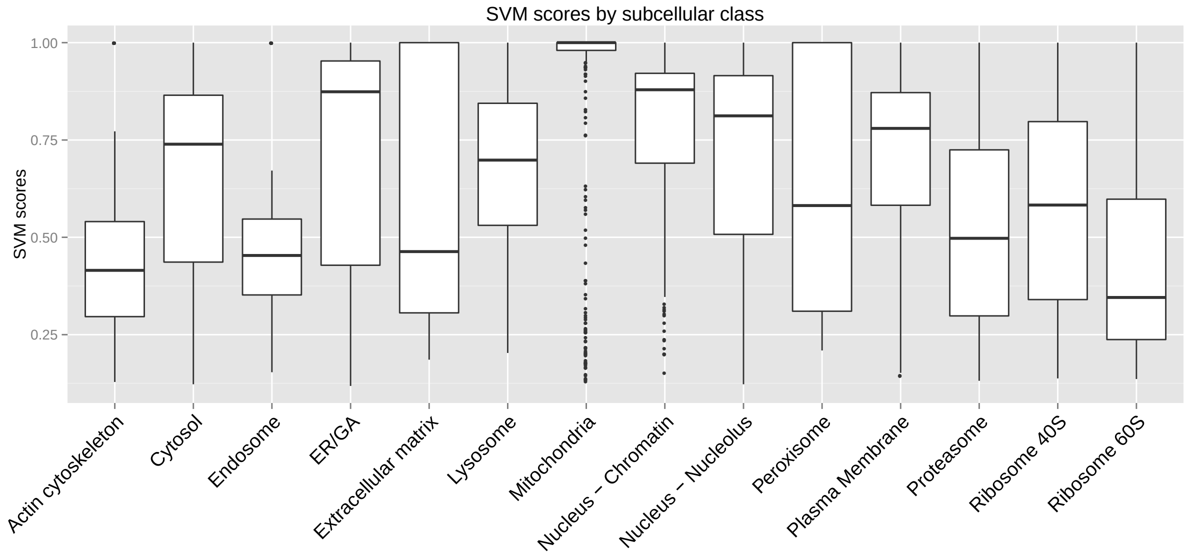
Supplementary Figure 13 | Spatial overview of the interaction partners of pluripotency triad Oct4, Nanog and Sox2. (A-C) HyperLOPIT reveals that while a majority of identified interaction partners of these three transcription factors are nuclear, some interaction partners were found with a variety of extranuclear distributions. (D) The shared interactors of the three transcription factors are predominantly localized to nuclear chromatin.[10] [11] [12]



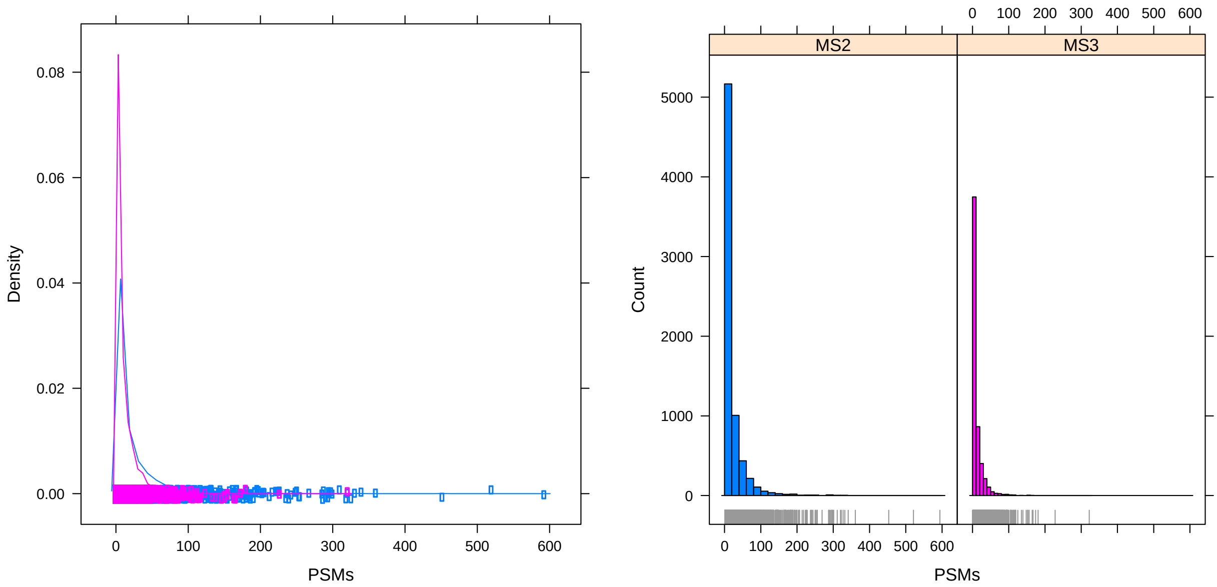
Supplementary Figure 14 | Examples of protein isoforms with differential subcellular localization in ES cells. (A) The long isoform of leucine aminopeptidase 3 (Lap3) was identified with mitochondrial localization, whereas the short isoform, which lacks the N-terminal import sequence, is localized between the cytosol and plasma membrane. Predicted interaction partners of Lap3 are found to localize across these three distributions, suggesting that the interactions are isoform-specific due to the differential compartmentalization of Lap3. (B) TMT 10-plex reporter ion profiles for the two isoforms of chromatin modifier Dnmt1 display differential localization. The long isoform enriches in the TMT 130C channel, consistent with chromatin localization, whereas the short isoform is most enriched in the TMT 129C channel, suggesting non-chromatin nuclear localization.



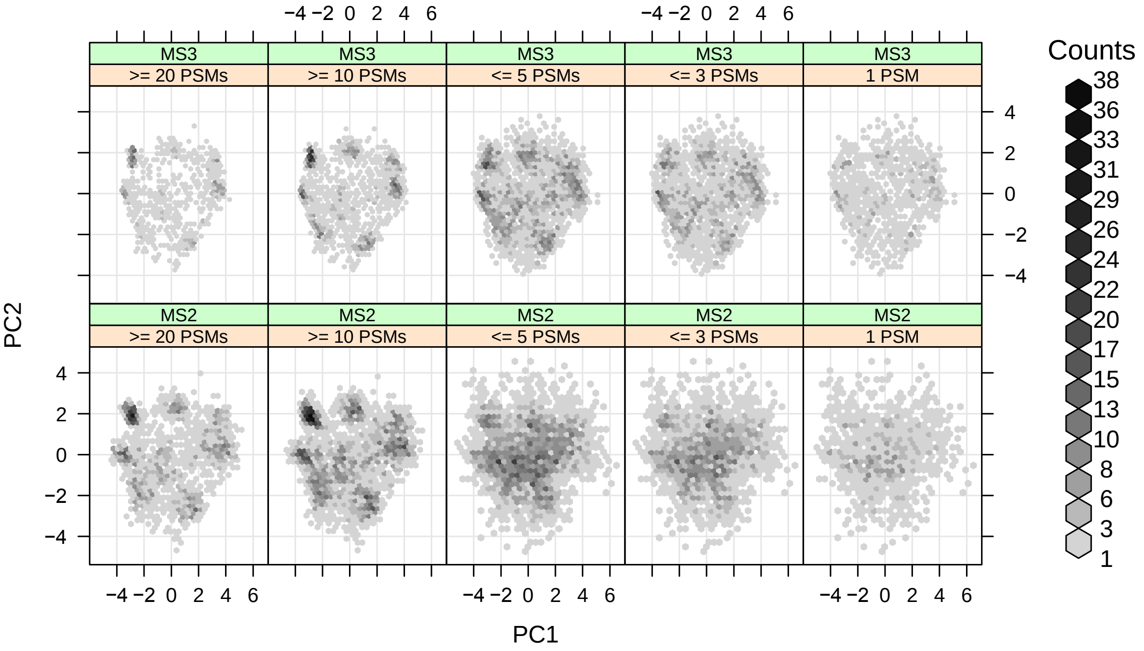
Supplementary Figure 15 | Application of novelty detection algorithm: initial marker proteins (top left), newly identified phenotypes (top right) and organelle markers after phenotype curation (bottom left).



Supplementary Figure 16 | Organelle specific SVM classification score distributions. Choosing a global single threshold is not satisfactory as different subcellular niches exhibit different score distributions reflective of their resolution.



Supplementary Figure 17 | Histograms (density, left and absolute counts, right) for the number of PSMs per protein for the MS^2 (blue) and SPS-MS^3 (magenta) data illustrating the higher number of proteins and the higher number of PSMs per proteins in MS^2.



Supplementary Figure 18 | PCA plot densities for MS^2 and $SPS\text{-}MS^3$.

MS Data Acquisition Mode	SPS precursors	Peptide-spectrum matches (PSMs)	PSMs with summed reporter ion intensity > 1 x 10^5	PSMs with summed reporter ion intensity > 1 x 10^6	Median Summed TMT 10-plex Reporter Intensity	Mean Missing TMT 10-plex Values
MS2	-	4197	97.50%	79.50%	3.3 x 10^6	0.2
MS3	-	4161	74.90%	43.90%	6.8 x 10^5	2.1
SPS-MS3	2	4126	81.90%	49.80%	9.8 x 10^5	1.5
	5	4173	88.20%	57.40%	1.5 x 10^6	1.0
	10	4192	92.80%	65.80%	2.1 x 10^6	0.6
	15	4195	94.50%	70.50%	2.5 x 10^6	0.5

Supplementary Table 1 | Comparison of ion signal and proportion of quantifiable PSMs for data acquired with conventional MS^2, conventional MS^3, and $SPS\text{-}MS^3$ with varying numbers of precursors. $SPS\text{-}MS^3$ with 10 precursors restores ion signal to levels comparable to MS^2, while improving the specificity and therefore accuracy of TMT quantification.

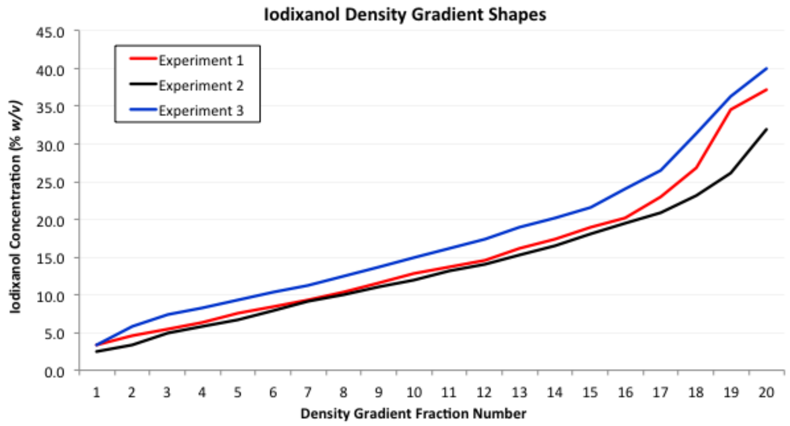
A

Fraction Number	Iodixanol Concentration (%)		
	Experiment 1	Experiment 2	Experiment 3
1	3.3	2.4	3.3
2	4.6	3.3	5.8
3	5.5	4.9	7.3
4	6.4	5.8	8.2
5	7.6	6.7	9.4
6	8.5	7.9	10.4
7	9.4	9.1	11.3
8	10.4	10.0	12.5
9	11.6	11.0	13.7
10	12.8	11.9	14.9
11	13.7	13.1	16.2
12	14.6	14.0	17.4
13	16.2	15.2	18.9
14	17.4	16.5	20.1
15	18.9	18.0	21.6
16	20.1	19.4	24.1
17	22.9	20.9	26.5
18	26.8	23.2	31.4
19	34.5	26.2	36.3
20	37.2	31.9	40.0
Cytosol*	0.0	0.0	0.0
Chromatin**	-	-	-

***Not collected from density gradient. Cytosol enriched fractions were collected from the supernatant of the crude membrane preparation step, as described in the Methods section.

****Not collected from density gradient. Chromatin enriched fractions were prepared using a parallel enrichment strategy based on detergent permeabilization, as described in the Methods section.

B



Supplementary Table 2 | Density gradient measurements (A) and plotted profiles (B) for three independent biological replicates.

TMT 10-plex Label	Density Gradient Fraction Number(s)			Subcellular Fraction Density (% *w/v* Iodixanol)		
	Experiment 1	**Experiment 2**	**Experiment 3**	**Experiment 1**	**Experiment 2**	**Experiment 3**
126	Cytosol	Cytosol	Cytosol	0.0	0.0	0.0
127N	1 to 6 (pooled)	1 to 6 (pooled)	1 to 4 (pooled)	6.0	5.2	6.2
127C	8 to 9 (pooled)	7 to 9 (pooled)	5 to 6 (pooled)	11.0	10.0	9.9
128N	10 to 11 (pooled)	10 to 11 (pooled)	8	13.3	12.5	12.5
128C	12	12	10	14.6	14.0	14.9
129N	14	14 to 15 (pooled)	12	17.4	17.3	17.4
129C	16	17	14	20.1	20.9	20.1
130N	18	18 to 19 (pooled)	17	26.8	24.7	26.5
130C	Chromatin	Chromatin	Chromatin	-	-	-
131	19	20	18	34.5	31.9	31.4

Supplementary Table 3 | Density gradient fractions selected for TMT 10-plex labeling in the three hyperLOPIT experiments.

Organelle	Number of proteins
ER/GA	76
Mitochondria	261
Plasma Membrane	50
Proteasome	34
Ribosome 40S	26
Ribosome 60S	43
unknown	3371

Supplementary Table 4 | Initial set of organelle marker proteins.

Organelle	Number of proteins
Phenotype 1	122
Phenotype 2	140
Phenotype 3	64
Phenotype 4	30
Phenotype 5	213
Phenotype 6	25
Phenotype 7	30
Phenotype 8	24
Phenotype 9	30
Phenotype 10	20
ER/GA	289
Mitochondria	449
Plasma Membrane	220
Proteasome	63
Ribosome 40S	63
Ribosome 60S	58
unknown	2021

Supplementary Table 5 | Assignments to novel phenotypes by the *phenoDisco* algorithm.

Organelle	Number of proteins
Actin cytoskeleton	13
Cytosol	43
Endosome	12
ER/GA	107
Lysosome	33
Mitochondria	383
Nucleus - Chromatin	64
Nucleus - Nucleolus	85
Peroxisome	17
Plasma Membrane	51
Proteasome	34
Ribosome 40S	27
Ribosome 60S	43
Unknown	2949

Supplementary Table 6 | Final augmented markers for SVM training.

Supplementary Note 1: Machine Learning Results

The first step of the classification process is to obtain a set of well-characterised organelle residents, termed protein 'markers'. These markers, once defined, can be used as input labelled data to train a machine learning classifier to assign proteins of unknown localisation to one of the localisations covered in the protein marker set. It is however laborious and extremely difficult to manually define reliable markers that cover the full sub-cellular diversity in the data, and furthermore to obtain markers that represent the true structure of any sub-cellular clusters determined, which is essential for sound analysis. As such, an initial round of phenotype discovery was conducted using the *phenoDisco* algorithm [13], in the *pRoloc* package [14].

Phenotype discovery

The *phenoDisco* algorithm uses iterative cluster merging combined with Gaussian Mixture Modelling and outlier detection, and with a minimal initial set of markers and unlabelled data can be used to effectively detect new putative clusters, beyond those that are initially manually described.

Ten new phenotype clusters were detected in the dataset (Supplementary Figure 15 and Supplementary Table 5). Each cluster was carefully validated by querying the UniProt database [15], the Gene Ontology [16] and the literature, as per the original pre-defined input markers, to assess biological relevance (Supplementary Table 6). Clusters that contained residents of small organellar structures such as the lysosome (phenotype 3) endosome (phenotype 4), and peroxisome (phenotype 7), were detected, thus confirming their independent data specific structure. Similarly, two very distinct nuclear clusters were confirmed, that were enriched in chromatin (phenotype 1) and nucleolus and other non-chromatin (phenotype 2) localised proteins. Further clusters contained actin cytoskeletal localised proteins (phenotype 9), ER localised proteins (phenotype 8) and a large cluster of cytosolic proteins (phenotype 5). We also see an interesting cluster that contains a small number of p-body proteins (phenotype 10) and a cluster of proteins that have mixed nuclear/cytoplasmic distributions (phenotype 6), of which many are known to shuttle between the nucleus and cytoplasm (see supplementary data 1 for *phenoDisco* output). Following examination of the phenotype clusters, further mining was conducted and well-known residents, as defined by UniProt and the literature, of the validated organelles were extracted and added to the list of protein markers to be used in a round of supervised machine learning classification. Markers for the lysosome, endosome, peroxisome, actin cytoskeleton, chromatin, nucleolus (non-chromatin) and cytosol were extracted from the discovery analysis to be added to the list of marker proteins. Proteins from phenotype 8, which are ER localised, were added to the existing set of ER markers, thus extending the number of markers for this organelle. Markers from phenotype 10 and phenotype 6 were left out of the final set of markers, as they were not highly enriched for one specific phenotype, and additionally the number of markers in these clusters was too small for use in classification (a minimum of 6 markers per subcellular class is required in supervised machine learning analysis for parameter optimisation as discussed in the proceeding section).

Increasing organellar resolution

Prior to novelty detection and supervised machine learning classification, to increase the organelle resolution, replicates 1 and 2 were combined using simple data fusion [17] in which quantitative TMT reporter ion ratios (10 per protein per experiment) were concatenated across the rows of proteins common in the two datasets. This combined dataset results in 20 quantitative data columns per protein and a total of 5032 proteins. Experiment 3 was not included as little additional resolution was obtained by further data fusion.

Comparison of MS^2 and SPS-MS^3 cluster resolution

Comparison of MS^2 and SPS-MS^3 protein-level cluster resolution and the repercussion for organelle proteomics has been investigated graphically as illustrated in Supporting Figures 17 and 18. The MS^2 and SPS-MS^3 (first replicate only) experiments contained 7116 and 5491 proteins respectively. Despite the higher number of proteins and peptide spectrum matches (PSMs) per proteins in MS^2, we demonstrate the negative impact of lack of accurate quantification on the sub-cellular resolution for proteins quantified by a limited number of PSMs. The histograms and density plots in Supplementary Figure 17, illustrate the higher number of proteins and PSM per protein in MS^2. Supplementary Figure 18 shows the SPS-MS^3 (top) and MS^2 (bottom) densities on the PCA plot for a set of PSM thresholds: from proteins with at least 20 PSMs per protein (left) to only a single PSM (right). Dense regions on the PCA plot are represented by darker shades on the figures. When considering proteins with a high number of PSMs (left), organelle clusters are clearly visible as darker groups. Filtering out proteins quantified by a high number of PMSs down to single PSM hits (right), the resolution of the sub-cellular clusters disappear already using a 5 PSM threshold in the MS^2 data; the density of point concentrates in the middle of the PCA figure, a pattern that characteristic of noisy, non-specific protein profiles. For SPS-MS^3 data, cluster resolution (organellar cluster densities and their separation) remains visible even for single PSM features.

Supplementary References

1 Bausch-Fluck, D. *et al.* A Mass Spectrometric-Derived Cell Surface Protein Atlas. *PLoS ONE* **10**, e0121314, doi:10.1371/journal.pone.0121314 (2015).

2 Braun, K. A. & Breeden, L. L. Nascent Transcription of MCM2-7 Is Important for Nuclear Localization of the Minichromosome Maintenance Complex in G(1). *Molecular Biology of the Cell* **18**, 1447-1456, doi:10.1091/mbc.E06-09-0792 (2007).

3 Betschinger, J. *et al.* Exit from Pluripotency Is Gated by Intracellular Redistribution of the bHLH Transcription Factor Tfe3. *Cell* **153**, 335-347 (2013).

4 Bouayad, D. *et al.* Nuclear-to-cytoplasmic Relocalization of the Proliferating Cell Nuclear Antigen (PCNA) during Differentiation Involves a Chromosome Region Maintenance 1 (CRM1)-dependent Export and Is a Prerequisite for PCNA Antiapoptotic Activity in Mature Neutrophils. *The Journal of Biological Chemistry* **287**, 33812-33825, doi:10.1074/jbc.M112.367839 (2012).

5 Krajewski, S. *et al.* Immunohistochemical analysis of Mcl-1 and Bcl-2 proteins in normal and neoplastic lymph nodes. *The American Journal of Pathology* **145**, 515-525 (1994).

6 Faou, P. & Hoogenraad, N. J. Tom34: A cytosolic cochaperone of the Hsp90/Hsp70 protein complex involved in mitochondrial protein import. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1823**, 348-357, doi:http://dx.doi.org/10.1016/j.bbamcr.2011.12.001 (2012).

7 Gegonne, A. *et al.* TFIID component TAF7 functionally interacts with both TFIIH and P-TEFb. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 5367-5372, doi:10.1073/pnas.0801637105 (2008).

8 Das, A. *et al.* RalA promotes a direct exocyst–Par6 interaction to regulate polarity in neuronal development. *Journal of Cell Science* **127**, 686-699, doi:10.1242/jcs.145037 (2014).

9 Havugimana, P. C. *et al.* A Census of Human Soluble Protein Complexes. *Cell* **150**, 1068-1081, doi:10.1016/j.cell.2012.08.011 (2012).

10 van den Berg, D. L. C. *et al.* An Oct4-Centered Protein Interaction Network in Embryonic Stem Cells. *Cell Stem Cell* **6**, 369-381, doi:10.1016/j.stem.2010.02.014 (2010).

11 Gagliardi, A. *et al.* A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *The EMBO Journal* **32**, 2231-2247, doi:10.1038/emboj.2013.161 (2013).

12 Gao, Z. *et al.* Determination of Protein Interactome of Transcription Factor Sox2 in Embryonic Stem Cells Engineered for Inducible Expression of Four Reprogramming Factors. *The Journal of Biological Chemistry* **287**, 11384-11397, doi:10.1074/jbc.M111.320143 (2012).

13 Breckels, L. M. *et al.* The effect of organelle discovery upon sub-cellular protein localisation. *Journal of Proteomics* **88**, 129-140 (2013).

14 Gatto, L., Breckels, L. M., Wieczorek, S., Burger, T. & Lilley, K. S. Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics*, doi:10.1093/bioinformatics/btu013 (2014).

15 The UniProt, C. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research* **40**, D71-D75, doi:10.1093/nar/gkr981 (2012).

16 The Reference Genome Group of the Gene Ontology, C. The Gene Ontology's Reference Genome Project: A Unified Framework for Functional Annotation across Species. *PLoS Computational Biology* **5**, e1000431, doi:10.1371/journal.pcbi.1000431 (2009).

17 Trotter, M. W. B., Sadowski, P. G., Dunkley, T. P. J., Groen, A. J. & Lilley, K. S. Improved sub-cellular resolution via simultaneous analysis of organelle proteomics data across varied experimental conditions. *Proteomics* **10**, 4213-4219 (2010).