Research

Towards defining the nuclear proteome

J Lynn Fink[¤*], Seetha Karunaratne[¤†], Amit Mittal[‡], Donald M Gardiner[†], Nicholas Hamilton[†§], Donna Mahony[†], Chikatoshi Kai[¶¥], Harukazu Suzuki[¶¥], Yosihide Hayashizaki[¶¥] and Rohan D Teasdale[*†]

Addresses: [*]ARC Centre of Excellence in Bioinformatics, The University of Queensland, St Lucia, Queensland, 4072, Australia. [†]Institute for Molecular Bioscience, University of Queensland, St Lucia, Queensland, 4072, Australia. [‡]Department of Biochemical Engineering and Biotechnology, Indian Institute of Technology, New Delhi, 110016, India. [§]Advanced Computational Modelling Centre, Department of Mathematics, University of Queensland, St Lucia, Queensland, 4072, Australia. [¶]Genome Exploration Research, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan. [¥]Genome Science Laboratory, Discovery Research Institute, RIKEN Wako Institute, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan.

[¤] These authors contributed equally to this work.

Correspondence: Rohan D Teasdale. Email: R.Teasdale@imb.uq.edu.au

Published: 23 January 2008

Genome ***Biology*** 2008, **9:**R15 (doi:10.1186/gb-2008-9-1-r15)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/1/R15

Received: 15 August 2007
Revised: 19 December 2007
Accepted: 23 January 2008

Abstract

Background: The nucleus is a complex cellular organelle and accurately defining its protein content is essential before any systematic characterization can be considered.

Results: We report direct evidence for 2,568 mammalian proteins within the nuclear proteome: the nuclear subcellular localization of 1,529 proteins based on a high-throughput subcellular localization protocol of full-length proteins and an additional 1,039 proteins for which clear experimental evidence is documented in published literature. This is direct evidence that the nuclear proteome consists of at least 14% of the entire proteome. This dataset was used to evaluate computational approaches designed to identify additional nuclear proteins.

Conclusion: This represents direct experimental evidence that the nuclear proteome consists of at least 14% of the entire proteome. This high-quality nuclear proteome dataset was used to evaluate computational approaches designed to identify additional nuclear proteins. Based on this analysis, researchers can determine the stringency and types of lines of evidence they consider to infer the size and complement of the nuclear proteome.

Background

Determination of organellar proteomes - the complement of proteins that reside, even if temporarily, in a specific organelle or subcellular region - is of fundamental importance. Cells are compartmentalized into membrane-bound structures in which specific biochemical processes occur and the function of these proteins is generally highly related to the function of the structure. Once the entire complement of proteins for an individual organelle has been defined, we can begin to systematically understand the molecular networks that control the biological processes occurring within that region.

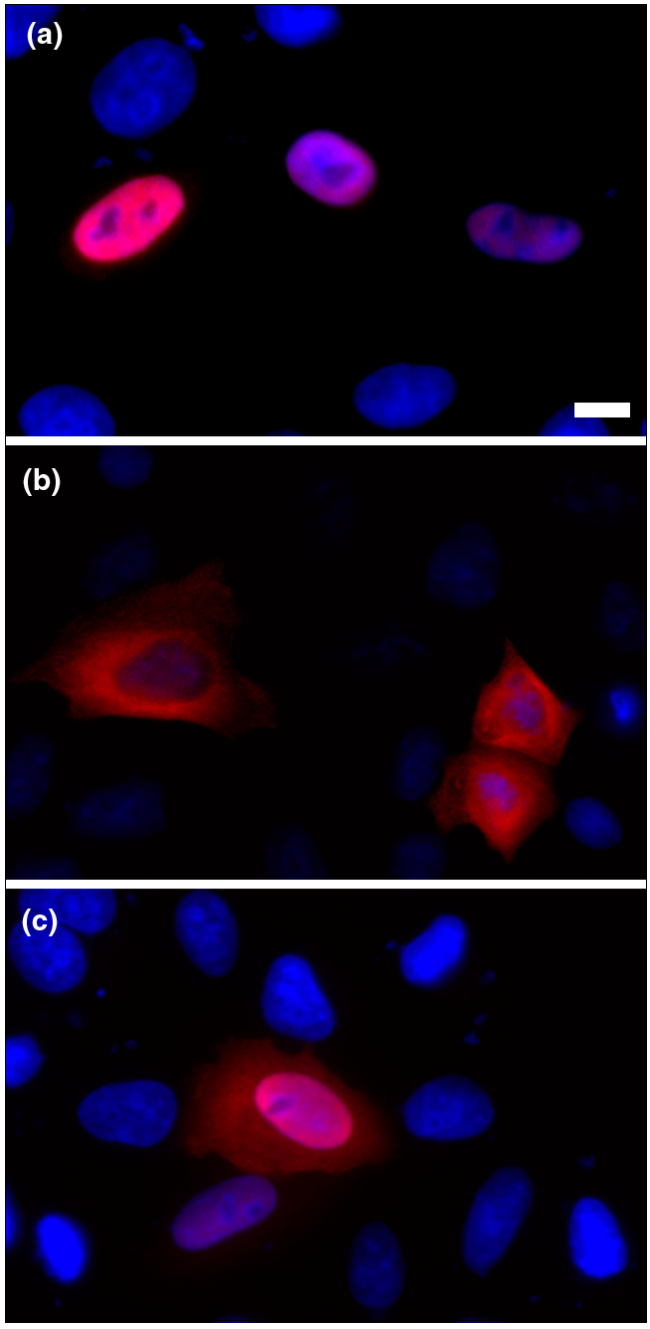

Figure 1
Representative immunofluorescence staining. Amino-terminal myc epitope-tagged expression constructs were generated and expressed in HeLa cells as described previously [15]. The scale bar represents 10 μm. **(a)** IκBα (U36277), a known nuclear protein, localizes to the nucleus. **(b)** Cyln2 (AAH53048) localizes to the cytoplasm. **(c)** Phf21b (AAH67021), a protein with no previous localization data, localizes to both the nucleus and cytoplasm.

The nucleus is an extremely complex organelle and is critical to the function of a eukaryotic cell. Therefore, identifying all of the proteins that localize to the nucleus is an important step towards understanding whole cell biology. Now that the genomes of several higher eukaryotes have been fully sequenced, this endeavor is beginning to become feasible. Recently, many high-throughput techniques have taken advantage of the availability of whole genomes and have been employed to localize organellar proteomes. These techniques include mass-spectrometry-based proteomics [1-3], genome-wide open-reading frame green fluorescent protein-tagging [4,5], and gene trap screens [6]. Here, we report the experimental subcellular localization of nuclear proteins in mouse using a high-throughput localization assay based on the expression of myc-tagged proteins. In addition, we extrapolate from this set to estimate the entire nuclear proteome using computational methods. Ultimately, the data presented here form the foundation for future studies into the functional aspects of nuclear biology such that the relationships and interactions between proteins and cellular processes can be explored in more detail.

Results and discussion

Subcellular localization assays identify 1,529 nuclear proteins

As a starting point for accurately defining the nuclear proteome, the proteins that were previously proposed to comprise the entire set of transcription regulators in mouse [7,8] were expressed as full-length, myc-tagged fusion proteins in HeLa cells and immunofluorescence was used to visualize each protein's subcellular location. This dataset of 1,559 was selected because transcription factors are known to act in the nucleus and should represent a large proportion of the nuclear proteome. In total, 1,253 proteins were assayed and localization data were captured for 1,056 proteins. There were 545 proteins that were observed in the nucleus only and 405 were observed in both the nucleus and the cytoplasm, resulting in a total of 950 nuclear proteins (Additional data file 1). Figure 1 shows images of proteins that localize to the nucleus, cytoplasm, or nucleus and cytoplasm and are representative of the images generated for all proteins. All image data have been warehoused in the LOCATE database and can be retrieved from LOCATE Subcellular Localization Database [9-11]. Fifteen proteins were observed to have subcellular localizations in addition to nuclear or cytoplasmic. This small subset predominantly represents inappropriate Gene Ontology (GO) annotations; for example, six are Rab GTPases and should not have been included in the original set of transcriptional regulators. For comparison, in our previous subcellular localization project directed at soluble phosphoregulators [12], only 40% of the proteins examined showed nuclear localization compared to 92% in this study. Importantly, within this study, 20 of 21 proteins reported in the literature to be cytoplasmic only were excluded from the nucleus [12].

It has been observed that organellar proteomes can vary between tissue types [13,14] so proteins that do not exhibit a nuclear localization in HeLa cells may localize to the nucleus in a different cell type. Therefore, the 106 proteins that were observed to reside only in the cytoplasm in HeLa cells were

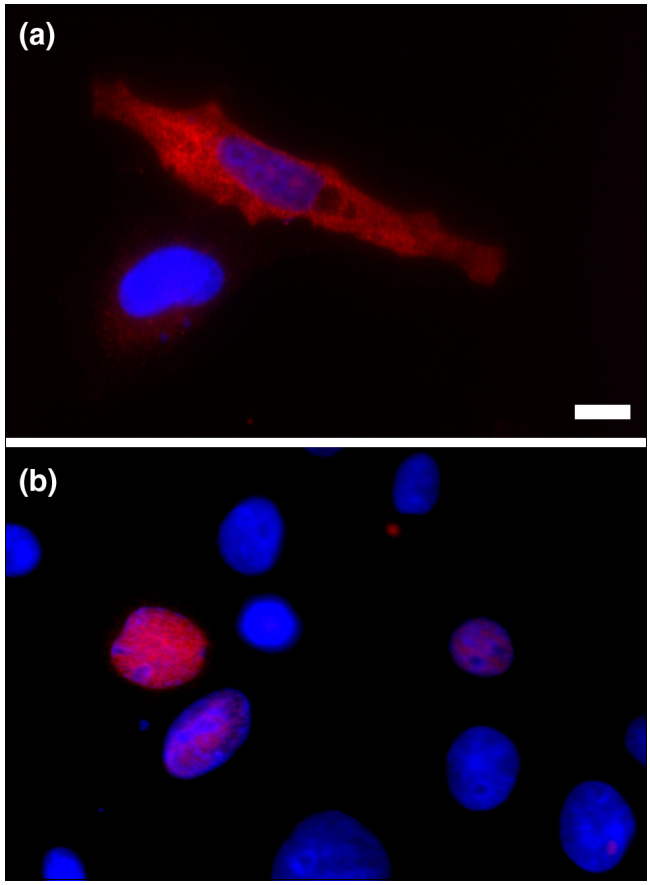

Figure 2
Expression of Trerf1 in two different cell lines. **(a)** Trerf1 (AAH59215) exhibits cytoplasmic localization in HeLa cells. **(b)** When Trerf1 is expressed in MCF7 cells, it localizes to the nucleus. The scale bar represents 10 μm.

expressed in MCF7 cells. Modulation of the cellular context revealed that 81 of these proteins were localized to the nucleus in MCF7 cells (Figure 2). The small number of 17 proteins, out of the 106 proteins successfully assayed, that did not display a nuclear subcellular localization could be members of the nuclear proteome that require an alternative condition or cellular context not considered in this study. For comparison we selected 200 proteins with a nucleus and cytoplasm subcellular localization in HeLa cells for expression in MCF7 cells. Only ten (five percent) of these proteins displayed a change in subcellular localization, with three classified as nuclear only and seven classified as cytoplasmic only. Eleven failed to yield a protein product that we were able to detect in MCF7 cells. These observations support that, for a fraction of proteins, the cellular context contributes significantly to their subcellular localization.

In addition to the transcriptional regulator set described above, we have applied the same approach to determine the subcellular localization of additional mouse proteins, including a set of putative type II membrane proteins previously reported [10,15]. As we generate the subcellular localization of these individual proteins the results are deposited within the LOCATE database. To date, we have experimentally observed an additional 498 proteins with a nuclear subcellular localization. Furthermore, within the LOCATE database, published data documenting a protein's subcellular localization are recorded with the original citation. According to the criteria outlined in Fink *et al*. [10], 1,247 proteins have a nuclear subcellular localization reported in the literature. Of the 1,529 nuclear proteins included in our experimental assay, 256 also have associated published literature reporting their subcellular localization, of which 80% of these are nuclear. Given that we have only documented direct evidence for 1,247 mammalian proteins to be part of the nuclear proteome, the identification of 1,529 novel nuclear proteins represents a significant increase in the number of proteins associated with this organelle (Figure 3). While the original transcriptional regulator set was biased towards transcription factors, only 60% of the final set of proteins experimentally localized are transcription factors (that is, have the GO nucleic acid binding annotation).

Other projects that document the subcellular localization of proteins based on direct evidence include the Nuclear Protein Database [16], which includes 1,227 proteins within the nuclear proteome based predominantly on literature and gene-trap experiments [6], and an alternative mammalian subcellular localization project [17] that reports 132 nuclear proteins in the LifeDB database [5,18]. UniProt v9.0 and MGI v3.5 (excluding proteins annotated as a result of reviewed computational analysis) contain evidence for 1,316 and 1,641 nuclear mouse proteins, respectively. However, a proportion of these proteins is not directly supported but is inferred based on a number of approaches.

We selected to exploit cell line models to define the nuclear proteome as they contain all the fundamental machinery to correctly target exogenously expressed proteins to the nucleus. Any tissue specific compartmentalization or regulation of subcellular localization will predominantly involve expression of additional binding proteins or regulatory pathways that function via post-translational modification, the majority of which are not likely to be functional within our subcellular localization assay. To determine if proteins examined in our subcellular localization assay with restricted expression are targeted differentially in our cell model we examined the GNF mouse tissue atlas data [19]. We considered a protein was 'restricted' if it was expressed in 1-40 tissue samples and 'broad' if expressed in 80-128 of the tissue samples. For the 553 proteins within the restricted class, 47% were nucleus only and 38% were nuclear/cytoplasmic, and within the 419 proteins broadly expressed, 38% were nucleus only and 47% were nuclear/cytoplasmic. These observations suggest that our subcellular localization assay captures a protein's intrinsic potential to localize into the nucleus regardless of its expression profile.

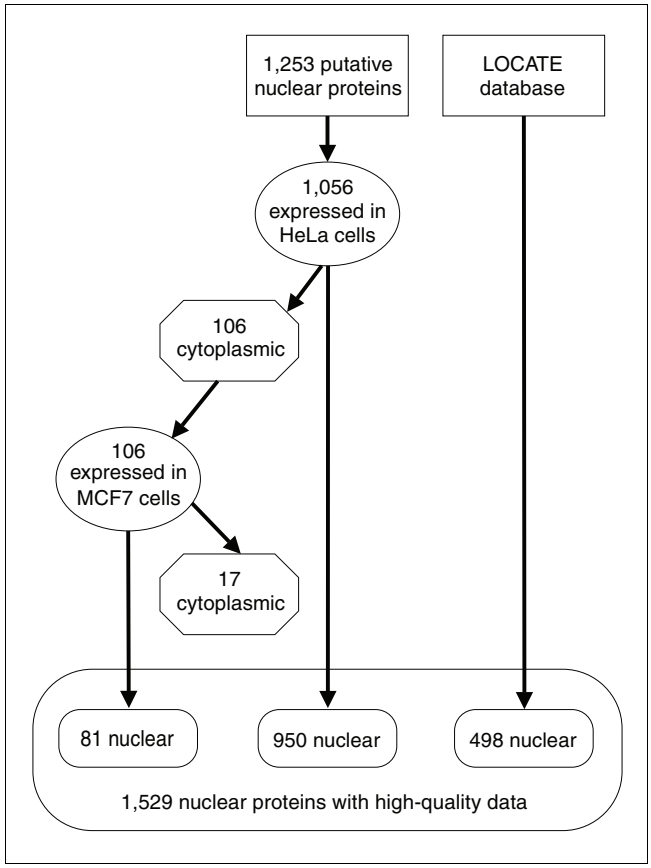

Figure 3
Flowchart describing experimental subcellular localization data acquisition. Experimental data were generated by expressing proteins in HeLa cells and determining their subcellular localization. Proteins that localized to the cytoplasm in HeLa cells were then expressed in MCF7 cells. Proteins reported to localize to the nucleus in the LOCATE database were also included in this dataset. Ultimately, all nuclear proteins were combined, resulting in a set of 1,529 proteins.

Combining the data from the subcellular localization assays with the set of nuclear proteins for which experimental evidence is documented in peer-reviewed literature, we created a high-quality nuclear proteome set of 2,568 proteins, termed NUCPROT. This dataset represents 14% of all mouse genes and contains protein products that have been experimentally confirmed to localize, at least in part, to the nucleus. Of particular importance are 532 proteins annotated as being a hypothetical protein [20] or as having an unknown function; the subcellular localization data reported here provide the first clues to their cellular roles.

Estimation of the size of the mouse nuclear proteome

Generation of the high-quality set of validated nuclear proteins, NUCPROT, enabled the critical evaluation of the distinct computational approaches developed to define the nuclear proteome. The following methods were used to predict additional putative nuclear proteins. Firstly, a number of computational methods have been developed as 'broad-based subcellular localization predictors' able to predict the subcellular localization of a protein to one or more of several specified locations. To further estimate the extent of the nuclear proteome, we selected five of these methods using criteria described previously [21] and applied them to the entire mouse proteome. We then extracted the proteins predicted to localize to the nucleus by each method. Secondly, it has been observed that the subcellular localization of proteins tends to be conserved across species and within protein families. 'Homology-based methods' to identify proteins related to the NUCPROT set will identify related proteins across species or within mouse. For example, a recent proteomics approach found that the human homolog to a yeast protein that localizes to the nucleolus has nearly a 90% chance of localizing to the same organelle [22]. Thirdly, another method of inferring nuclear localization of a protein is the prediction of nuclear localization signals (NLS). The NLS is a short peptide sequence that functions as a sorting signal to facilitate the import of a protein into the nucleus.

Details of the computational approaches applied are described in Materials and methods. Table 1 summarizes the results of each of these approaches applied to the high-quality set of nuclear proteins, NUCPROT, and also to the entire mouse proteome as defined [10,11,20]. Figure 4 shows the frequency with which each protein was classified as nuclear by these methods.

Using the homology approach, we compared mouse proteins that are homologous to yeast proteins that were determined to have a nuclear localization by a proteome-wide analysis of protein localization [4]. Using a stringent approach, we found 691 mouse proteins not already included in NUCPROT while a more permissive approach found 2,031 proteins. We also employed the homology approach to select other proteins within the mouse proteome that may be nuclear by inferring homologs to the NUCPROT dataset and found 766 additional homologous mouse proteins. Using the subcellular localization predictors, we found that between 4,084 and 9,122 proteins were predicted to localize to the nucleus and the NLS predictor, Nucleo, predicted that 987 proteins contain a signal.

Each of these methods has its own unique aspect and will likely define part of the nuclear proteome the others will fail to detect. Application of the inferred nuclear proteome will vary, so inclusion based on subsets of the nine computational approaches will be required. Sixty-two percent of the entire mouse proteome, including the NUCPROT set, was predicted by at least one method to be nuclear, resulting in a maximal nuclear proteome. However, a reasonable conservative estimate would include those proteins that were predicted by four out of nine methods, a threshold that included over half of the NUCPROT set. This results in a nuclear proteome of 28% of the mouse proteome, or 5,422 proteins. We term this set of proteins the NUCPROT+4-inferred dataset.

Table 1

Results from computational approaches predicting nuclear proteome membership

Method	NUCPROT proteins classified as 'nuclear'	Accuracy	RIKEN proteome proteins classified as 'nuclear'
CELLO	2,125	78%	9,122 (47%)
pTARGET	1,706	63%	5,953 (30%)
Proteome Analyst	1,803	66%	4,084 (21%)
WoLF PSORT	1,909	70%	7,172 (37%)
MultiLoc	1,561	57%	5,137 (26%)
Yeast homology ($E < 10^{-4}$)	218	8.0%	2,031 (8.0%)
Yeast homology ($E < 10^{-30}$)	47	1.7%	691 (3.5%)
Mouse homology	430	16%	766 (3.9%)
Nucleo	857	32%	987 (5.0%)

For the subcellular localization prediction programs, proteins were considered to be incorrectly classified as 'not nuclear' if the method's top-ranked localization call was not 'nucleus' but the protein was in our high-quality dataset; proteins were considered to be correctly classified as 'nuclear' if the method's top-ranked localization call was 'nuclear' and the protein was in our high-quality dataset.

We can compare our NUCPROT+4-inferred dataset to other nuclear proteome estimates. Firstly, based on the combination of any subcellular localization annotation from all major protein databases, 3,848 mouse proteins have been annotated as nuclear within the subcellular localization database LOCATE [10,11]. While these annotations are frequently

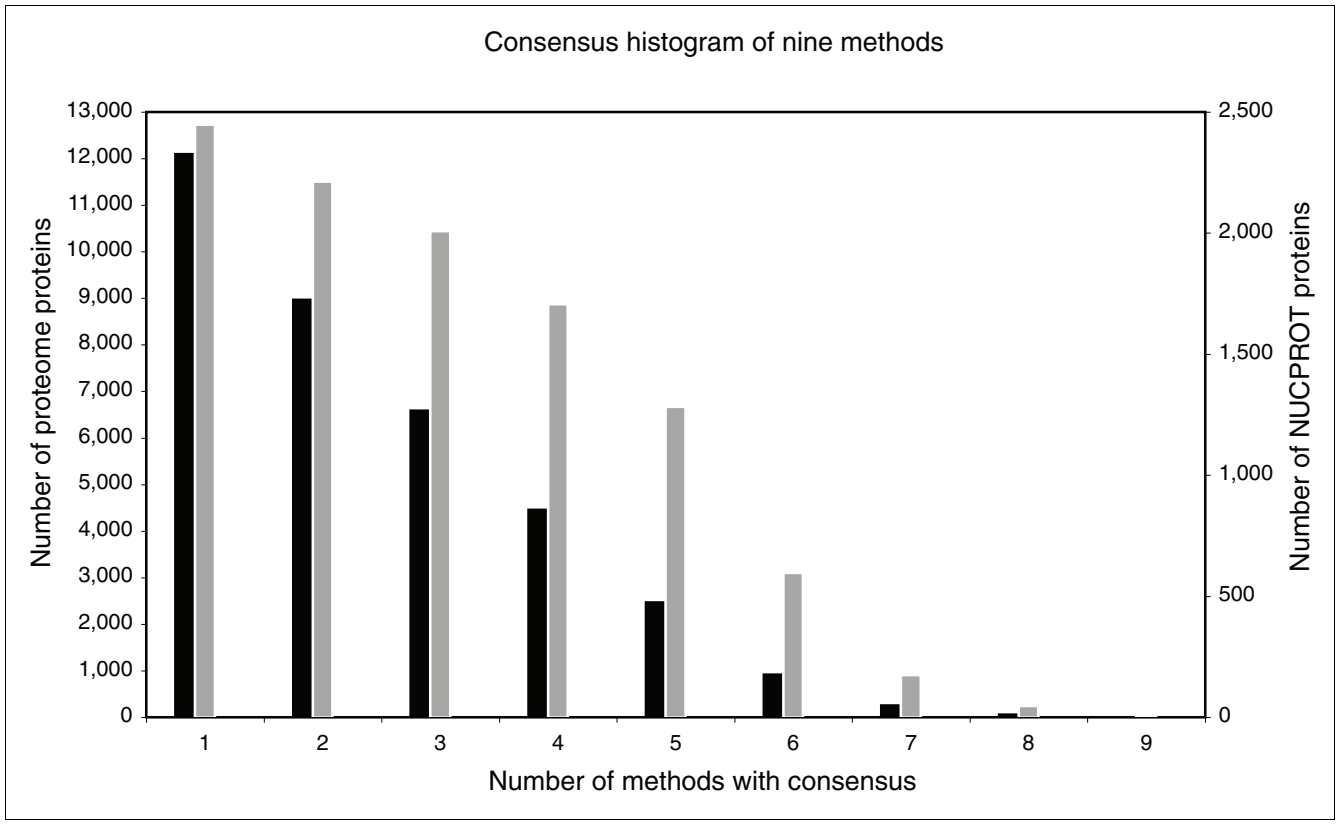

Figure 4
Consensus histogram of nine localization methods. The numbers of proteins that were predicted to be nuclear by each of nine methods are shown as bars. We selected proteins that were predicted to be nuclear by at least four methods. The black bars represent the proteins from the entire mouse proteome while the gray bars represent proteins from the NUCPROT set.

not supported by direct experimental evidence, they represent an estimate of the current nuclear proteome. For comparison, 91% of these proteins are included in the NUCPROT+4-inferred dataset. Secondly, of the 5,422 NUCPROT+4-inferred dataset, only 49% of these are annotated with the 'cellular component' GO term 'nucleus' in the MGI database.

Conclusion

Obtaining a fully defined nuclear proteome with each protein having been experimentally localized to the nucleus will ultimately require the determination of the subcellular localization of the entire mammalian proteome and, based on our observations, across a range of cell types. At the moment this is beyond peptide sequencing proteomic approaches and will likely require high-throughput epitope-tagged cell expression approaches like that commenced in this study. Our initial survey puts us on the path towards defining the nuclear proteome with direct evidence that 14% of the mouse proteome contributes to the nuclear proteome. Having defined individual proteins as contributing to the nuclear proteome, then clearly the next stage is to delineate the nucleocytoplasmic trafficking pathways [23] that contribute to each individual nuclear protein's distribution and how their subcellular localization is regulated under distinct cellular conditions.

Materials and methods

Dataset

The mouse proteome dataset selected for this analysis was the Isoform Protein Sequence set created by the RIKEN FANTOM3 Consortium from novel and public protein coding transcripts and has been described previously [7,10,24]. This dataset was supplemented with additional sequences reported to belong to the set of mouse transcriptional regulators [8] and consists of a total of 19,562 transcriptional units. Within the text we do not consider the multiple isoforms generated from a single protein coding transcriptional unit.

Protein subcellular localization

The methods published by Aturaliaya *et al.* [15] were followed, unless stated otherwise, for making expression constructs, transfection into cells, and immunolabeling and image capture. Expression constructs of cDNA clones with an amino-terminal myc epitope were generated and transfected into HeLa and/or MCF7 cells cultured at 60-70% confluency. Expression of proteins was detected 24 hours after transfection by immunolabeling with monoclonal anti-myc antibody (Cell Signaling Technology, Inc., Boston, MA, USA). Cell monolayers were treated with DAPI to label the nuclei. Images were captured on an Olympus AX-70 upright fluorescence microscope. Protein localization data were classified into 'nuclear', 'cytoplasmic', and 'nuclear and cytoplasmic' based on the predominant type of expression in each transfected sample.

Automated image classification

Subcellular localizations were inferred from the images by both an expert curator and the automated image classification program ASPiC [25]. ASPiC is a fully automated system that assigns a subcellular location to an image. It selects, masks and crops cells within each image, using a corresponding DAPI image to localize the nucleus, generates image statistics, and produces an automated classification for each cropped cell image using a support vector machine. If, for a given protein, there are multiple cells with multiple classifications, a vote is taken to give an overall classification. Average image intensities and areas of the nuclear and non-nuclear regions are also recorded for each cropped cell. Three out of 1,608 images classified by ASPiC were assigned locations that conflicted with the location assigned by a human curator; these conflicts were resolved during a manual review by a second expert curator.

Computational predictions

Predictions using programs that predict subcellular localization to multiple cellular locations were performed as described previously [21]. Briefly, publicly available programs that predicted localization to at least nine major locations (nucleus, cytoplasm, mitochondrion, extracellular region, plasma membrane, Golgi apparatus, endoplasmic reticulum, peroxisome, and lysosome) and could accept large sequence batches were used to predict locations for all proteins encoded by the mouse transcriptome; these were CELLO [26], WoLF PSORT [27], MultiLOC [28], Proteome Analyst [29,30], and pTARGET [31].

Nuclear localization signals were predicted by predictNLS [32,33], NucPred [34], and Nucleo [35]. NucPred and Nucleo predictions at or above 0.8 were considered to be positive.

Homology inference

Homologs were inferred by performing a BLAST search [36] of the entire mouse proteome with itself and with nuclear yeast proteins from the Yeast GFP Fusion Localization Database [4]. BLAST hits that did not have sequence coverage of 50% or more were discarded from further analysis. An optimal E-value threshold for selecting homologs was determined by maximizing the number of positives while minimizing the number of negatives using the set of high-confidence nuclear mouse proteins as a set of true positives and the remainder of the mouse proteome as a set of true negatives. The optimal E-value threshold was 10^{-140} for mouse proteins and 10^{-4} for yeast proteins. An additional, more stringent E-value threshold of 10^{-30} was selected for the yeast proteins based on a previous study of computed gene homology [37].

Abbreviations

GO, Gene Ontology; NLS, nuclear localization signals.

Authors' contributions

SK, AM, DG and DM participated in the generation of experimental cell biology data. JLF performed the bioinformatics studies, integration of the data and drafted the manuscript. NH developed and implemented the image analysis protocols. CK, HS and YH designed and generated all the transcript templates used in this study. RDT conceived of the study, and participated in its design and coordination. All authors read and approved the final manuscript.

Additional data files

The following additional data are available. Additional data file 1 is a table listing the subcellular localization data generated in this project and the nuclear proteome, NUCPROT+4-inferred. The table includes a list of proteins in the nuclear proteome, the experimentally determined subcellular location, and data annotating the degree of confidence in their membership. Proteins are referred to by Entrez Gene_ID, GenBank accession number, and b RIKEN representative protein set ID [20].

Acknowledgements

The authors would like to thank Kelly Hanson for assistance with the subcellular localization assays and John Hawkins and Mikael Bodén for their assistance with Nucleo. This work was supported by funds from the following: Australian Research Council of Australia; Australian National Health and Medical Research Council of Australia; Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to YH; a grant of the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan; and a Grant for the RIKEN Frontier Research System, Functional RNA research program. RDT is supported by an NHMRC R Douglas Wright Career Development award. NH is partially supported by the Australian Research Council's award of a Federation Fellowship to Prof. Kevin Burrage.

References

1. Roix J, Misteli T: **Genomes, proteomes, and dynamic networks in the cell nucleus.** *Histochem Cell Biol* 2002, **118:**105-116.
2. Simpson JC, Pepperkok R: **Localizing the proteome.** *Genome Biol* 2003, **4:**240.
3. Simpson JC, Pepperkok R: **The subcellular localization of the mammalian proteome comes a fraction closer.** *Genome Biol* 2006, **7:**222.
4. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425:**686-691.
5. Mehrle A, Rosenfelder H, Schupp I, del Val C, Arlt D, Hahne F, Bechtel S, Simpson J, Hofmann O, Hide W, Glatting KH, Huber W, Pepperkok R, Poustka A, Wiemann S: **The LIFEdb database in 2006.** *Nucleic Acids Res* 2006, 34(Database issue):D415-D418.
6. Sutherland HG, Mumford GK, Newton K, Ford LV, Farrall R, Dellaire G, Cáceres JF, Bickmore WA: **Large-scale identification of mammalian proteins localized to nuclear sub-compartments.** *Hum Mol Genet* 2001, **10:**1995-2011.
7. Kanamori M, Konno H, Osato N, Kawai J, Hayashizaki Y, Suzuki H: **A genome-wide and nonredundant mouse transcription factor database.** *Biochem Biophys Res Commun* 2004, **322:**787-793.
8. Nilsson R, Bajic VB, Suzuki H, di Bernardo D, Bjorkegren J, Katayama S, Reid JF, Sweet MJ, Gariboldi M, Carninci P, Hayashizaki Y, Hume DA, Tegner J, Ravasi T: **Transcriptional network dynamics in macrophage activation.** *Genomics* 2006, **88:**133-142.
9. **LOCATE Subcellular Localization Database** [http://locate.imb.uq.edu.au]
10. Fink JL, Aturaliya RN, Davis MJ, Zhang F, Hanson K, Teasdale MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Teasdale RD: **LOCATE: a mouse protein subcellular localization database.** *Nucleic Acids Res* 2006, 34(Database issue):D213-D217.
11. Sprenger J, Lynn Fink J, Karunaratne S, Hanson K, Hamilton NA, Teasdale RD: **LOCATE: a mammalian protein subcellular localization database.** *Nucleic Acids Res* 2008, 36(Database issue):D230-D233.
12. Forrest AR, Taylor DF, Fink JL, Gongora MM, Flegg C, Teasdale RD, Suzuki H, Kanamori M, Kai C, Hayashizaki Y, Grimmond SM: **PhosphoregDB: the tissue and sub-cellular distribution of mammalian protein kinases and phosphatases.** *BMC Bioinformatics* 2006, **7:**82.
13. Mootha VK, Bunkenborg J, Olsen JV, Hjerrild M, Wisniewski JR, Stahl E, Bolouri MS, Ray HN, Sihag S, Kamal M, Patterson N, Lander ES, Mann M: **Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria.** *Cell* 2003, **115:**629-640.
14. Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, Scott MS, Gramolini AO, Morris Q, Hallett MT, Rossant J, Hughes TR, Frey B, Emili A: **Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling.** *Cell* 2006, **125:**173-186.
15. Aturaliya RN, Fink JL, Davis MJ, Teasdale MS, Hanson KA, Miranda KC, Forrest AR, Grimmond SM, Suzuki H, Kanamori M, Kai C, Kawai J, Carninci P, Hayashizaki Y, Teasdale RD: **Subcellular localization of mammalian type II membrane proteins.** *Traffic* 2006, **7:**613-625.
16. Dellaire G, Farrall R, Bickmore WA: **The Nuclear Protein Database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome.** *Nucleic Acids Res* 2003, **31:**328-330.
17. Simpson JC, Wellenreuther R, Poustka A, Pepperkok R, Wiemann S: **Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing.** *EMBO Rep* 2000, **1:**287-292.
18. Bannasch D, Mehrle A, Glatting KH, Pepperkok R, Poustka A, Wiemann S: **LIFEdb: a database for functional genomics experiments integrating information from external sources, and serving as a sample tracking system.** *Nucleic Acids Res* 2004, 32(Database issue):D505-D508.
19. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101:**6062-6067.
20. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, *et al.*: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309:**1559-1563.
21. Sprenger J, Fink JL, Teasdale RD: **Evaluation and comparison of mammalian subcellular localization prediction methods.** *BMC Bioinformatics* 2006, **7(Suppl 5):**S3.
22. Andersen JS, Lam YW, Leung AK, Ong SE, Lyon CE, Lamond AI, Mann M: **Nucleolar proteome dynamics.** *Nature* 2005, **433:**77-83.
23. Terry LJ, Shows EB, Wente SR: **Crossing the nuclear envelope: hierarchical regulation of nucleocytoplasmic transport.** *Science* 2007, **318:**1412-1416.
24. Davis MJ, Hanson KA, Clark F, Fink JL, Zhang F, Kasukawa T, Kai C, Kawai J, Carninci P, Hayashizaki Y, Teasdale RD: **Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units.** *PLoS Genet* 2006, **2:**e46.
25. Hamilton NPR, Hanson K, Fink L, Karunaratne S, Teasdale RD: **Automated subcellular phenotype classification.** In *Conferences in Research and the Practice in Information Technology Volume 73*. Australian Computer Society; 2006.
26. Yu CS, Chen YC, Lu CH, Hwang JK: **Prediction of protein subcellular localization.** *Proteins* 2006, **64:**643-651.
27. Horton P: **Protein subcellular localization prediction with WoLF PSORT.** In *Fourth Asia-Pacific Bioinformatics Conference: February 13-16 2006; Taipei* Edited by: Jiang T, Yang UC, Chen YPP, Wong L. London: Imperial College Press; 2006:39-48.
28. Höglund A, Dönnes P, Blum T, Adolph HW, Kohlbacher O: **MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition.** *Bioinformatics* 2006, **22:**1158-1165.

29. Szafron D, Lu P, Greiner R, Wishart DS, Poulin B, Eisner R, Lu Z, Anvik J, Macdonell C, Fyshe A, Meeuwis D: **Proteome Analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations.** *Nucleic Acids Res* 2004, 32(Web Server issue):W365-W371.
30. Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, Poulin B, Anvik J, Macdonell C, Eisner R: **Predicting subcellular localization of proteins using machine-learned classifiers.** *Bioinformatics* 2004, **20:**547-556.
31. Guda C: **pTARGET: a web server for predicting protein subcellular localization.** *Nucleic Acids Res*2006, 34(Web Server issue):W210-W213.
32. Nair R, Carter P, Rost B: **NLSdb: database of nuclear localization signals.** *Nucleic Acids Res* 2003, **31:**397-399.
33. Cokol M, Nair R, Rost B: **Finding nuclear localization signals.** *EMBO Rep* 2000, **1:**411-415.
34. Heddad A, Brameier M, MacCallum RM: **Evolving regular expression-based sequence classifiers for protein nuclear localisation.** *Lecture Notes Computer Sci* 2004, **3005:**31-40.
35. Hawkins J, Davis L, Boden M: **Predicting nuclear localization.** *J Proteome Res* 2007, **6:**1402-1409.
36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
37. Gilbert DG: **euGenes: a eukaryotic genome information system.** *Nucleic Acids Res* 2002, **30:**145-148.