MINORITY REPORT: FUNCTIONAL CHARACTERIZATION OF A SUBSET OF EXTREMELY ALKALINE PROTEINS IN THE HUMAN PROTEOME

Alexander A. Tokmakov[1] and Atsushi Kurotani[2]

Abstract

Whole-proteome distributions of the protein isoelectric point (pI) are not Gaussian but rather multimodal. In the present study, we performed functional characterization of an extremely alkaline protein modality (pI>11.5, n=503) in the human proteome. Here, we report that the majority of the extremely alkaline proteins (69%), which have predominantly nuclear or mitochondrial localization (80%), are involved in information storage and processing. These proteins were further classified in the functional groups of "Translation, ribosomal structure and biogenesis", "RNA processing and modification", and "Chromatin structure and dynamic", according to the classification of Eukaryotic Orthologous Groups, KOGs. In addition, a lot of nuclear proteins in the analyzed subset (63%) bear the nucleolar localization signal. These data indicate that many of the extremely alkaline proteins localize to the nucleolus and are involved in ribosome biogenesis and RNA processing.

Key words: protein pI, human proteome, multimodality, extremely alkaline proteins, protein function

Introduction

Proteome-wide distributions of the protein isoelectric point (pI) in different organisms are multimodal with two major acidic and alkaline peaks and several minor sub-peaks and shoulders (Schwartz et al. (1), Knight et al. (2), Wu et al. (3), Carugo (4)). Accordingly, our recent bioinformatics studies revealed that the whole-proteome distribution of protein pI in the human proteome is essentially bimodal with some minor statistical features (Kurotani et al. (5), Tokmakov et al. (6), Fig. 1A). It was previously demonstrated that the two major modes in the whole-proteome pI distributions correspond to the pI distributions of cytosolic and integral membrane proteins (Schwartz et al. (1), Knight et al. (2)), however the minor observed features have not been associated with specific protein subsets (Wu et al. (3), Carugo (4)).

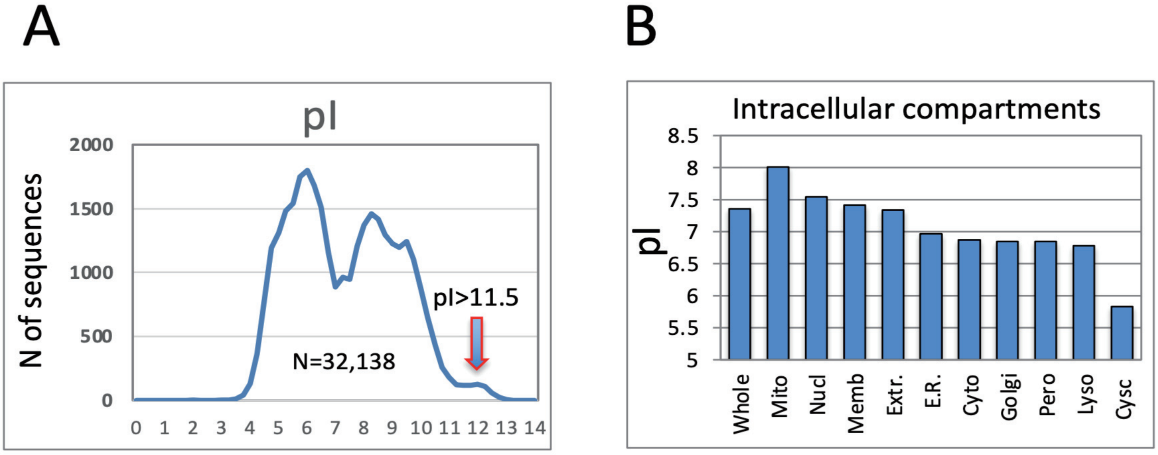

Fig 1. Distributions of protein pIs in the human proteome. The distribution of pI values calculated for 32,138 predicted human proteins is shown in panel (**A**). The modality of extremely alkaline proteins is denoted by an arrow. Panel (**B**) presents mean pI values of the protein pI distributions in different subcellular compartments.

Received 23 December 2021

[1] Institute of Advanced Technology, Faculty of Biology-Oriented Science and Technology, KinDai University, 930 Nishimitani, Kinokawa city, Wakayama 649-6493, Japan

[2] Center for Sustainable Resource Science, RIKEN Yokohama Institute, 1-7-22, Suehiocho, Yokohama, Kanagawa 230-0045, Japan

Importantly, it was found that protein pI distributions and their averaged values differ significantly in subcellular compartments (Ho et al. [7], Brett et al. [8], Kiraga et al. [9]). The studies converged on the assumption that subcellular localization-specific patterns of protein pI are defined by local environmental constraints. Also, our recent investigation using one of the latest updates of human genome data demonstrated that protein pI correlates positively with nuclear and mitochondrial localizations and negatively with cytoskeletal, lysosomal, peroxidase, Golgi and cytoplasmic ones (Kurotani et al. [5], Fig. 1B). The study also revealed that organelle-specific pI distributions are defined by local pH and membrane charge.

One of the common tasks of big data mining is anomaly/outlier detection and identification of unusual data records that require further investigation. In the present study, we carried out functional characterization of a minor subset of extremely alkaline proteins (pI>11.5, n=503) observed in the human proteome (Fig. 1A). It was demonstrated previously that the proteins of nuclear and mitochondrial localizations are overrepresented in this subset (Kurotani et al. [5], Fig. 2). Although the minor peak of extremely alkaline proteins was distinguished in pI distributions of some eukaryotic proteomes (Wu et al. [3], Carugo [4]), it has not been explicitly characterized so far. In particular, the very existence of the extremely alkaline proteins raises concern about their biological function at the molecular and cellular levels.

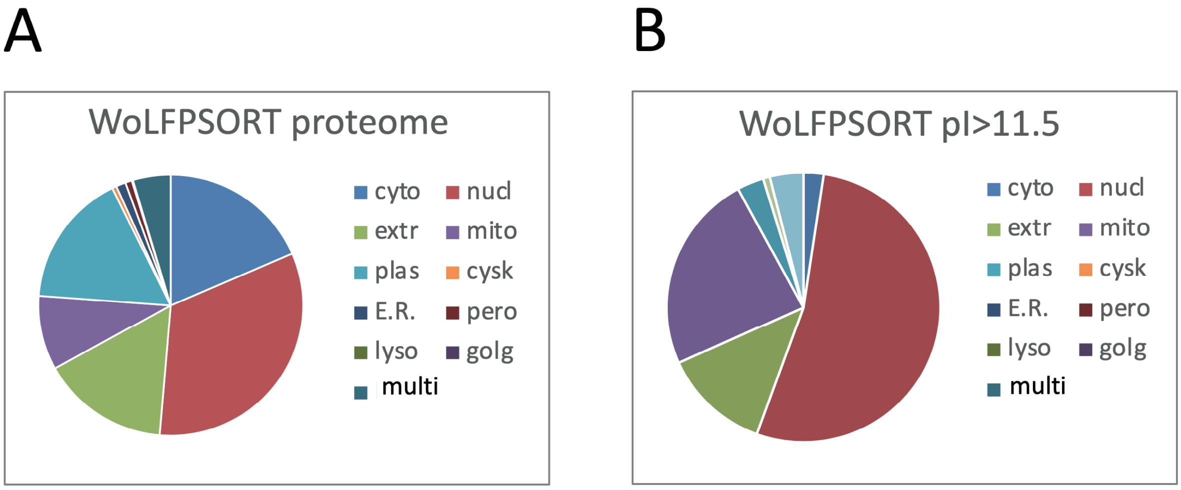

Fig 2. Assignments of protein subcellular localization for the studied datasets. Relative contents of the proteins at ten subcellular locations and the proteins predicted to locate in multiple compartments (denoted as "multi") were determined for the complete dataset (**A**) and the subset of extremely alkaline proteins (**B**) of the human proteome. Intracellular localization of the individual proteins was predicted using the WoLF PSORT algorithm.

Materials and methods

Data sets

The complete human proteome dataset was constructed using the proteome resource available at ftp://ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens/protein/. The redundancy check was carried out using the CD-HIT tool (Fu et al. [10]) to remove amino acid sequences with more than 90% identity. The sequences containing less than 50 amino acids were also filtered out. The total number of sequences in the final whole-proteome dataset was 32,138. A subset of extremely alkaline human proteins was extracted from the complete human proteome dataset. It comprised 503 amino acid sequences with calculated pI values > 11.5, covering about 1.5% (503/32,138) of the whole proteome. Markedly, this subset did not contain histones, because pIs of these highly alkaline nuclear proteins are lower than the threshold level set for the dataset. The pI values for the core histones, such as H2A and H2B and their variants, are in the interval of 10.0-10.5, and those for H3 and H4 fall in the range of 11.0-11.5. The human linker histone H1 has pI=11.13.

Calculation and prediction of protein properties

Protein pI values were calculated using the free Prot-Param tool (Gasteiger et al. [11]) provided at the ExPASy server (https://web.expasy.org/protparam/). Protein localization was predicted with the WoLF PSORT Advanced Protein Subcellular Localization Prediction Tool (Horton et al. [12]), downloadable from the GenScript server (https://www.genscript.com/wolf-psort.html). Protein annotations of the KOG, EuKaryotic Orthologous Groups, was assigned using the BLASTP program with an e-value lower than 1e-10. The KOG database is provided at the NCBI site (https://www.ncbi.nlm.nih.gov/COG) (Koonin et al. [13]). Nucleolar localization signals (NoLS) in the amino acid sequences of extra-alkaline subset were identified with the NoD, Nucleolar Localization Sequence Detector, predictive tool available online (http://www.compbio.dundee.ac.uk/www-nod/) (Scott et al. [14]).

Results and discussion

Functional assignments of extremely alkaline human proteins

In contrast to numerous investigations of the relationships between protein pI and subcellular localization, no comprehensive proteome-wide study addressing specifically correlations between protein pI and function has been presented so far. The existence of such correlations can be suggested by the fact that the proteins and functional domains involved in the same cellular function often display similar physicochemical and structural properties. Moreover, the proteins of differing functional classes should necessarily be employed at specific subcellular locations, considering divergent functions of intracellular organelles.

It was demonstrated previously that mitochondrial and nuclear proteins are predominant in the subset of extremely alkaline proteins in stark contrast to the whole proteome (Kurotani et al. [5], Fig. 2). As it could be expected, general functional assignments of proteins in the two datasets differed too. The proteins of “Information storage and processing” were profoundly abundant, whereas the proteins in the categories of “Cellular processes and signaling” and “Metabolism” were largely underrepresented in the alkaline subset (Fig. 3).

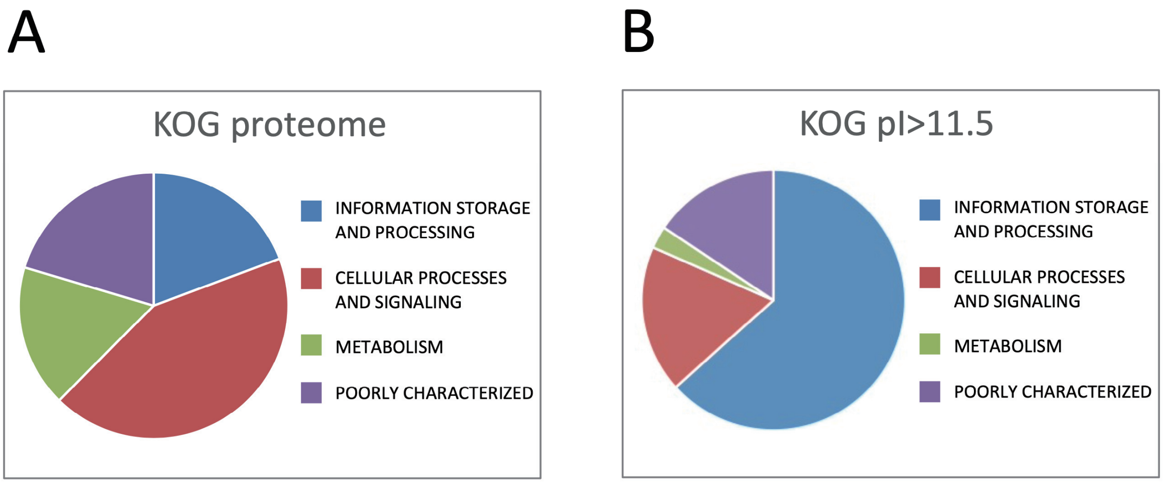

Fig 3. Assignments of general protein function for the studied subsets. Functional assignments were carried out for the complete dataset (**A**) and the subset of extremely alkaline proteins (**B**) of the human proteome. General functional categories of the individual proteins were assigned using the KOG classification algorithm.

More detailed assignment of protein functions using the KOG classificator revealed that the proteins of “Information storage and processing” in the extremely alkaline subset were mainly represented by certain functional categories, such as “Translation, ribosomal structure and biogenesis” (J), “RNA processing and modification” (A), and “Chromatin structure and dynamic” (B) (Fig. 4A,B). The proteins of these functions were largely overrepresented (5 to 10-fold) in the alkaline subset, as compared to the whole proteome dataset (Fig. 4C). On the other hand, the most abundant functional categories of “Cellular processes and signaling”, such as “Signal transduction mechanisms” (T), “Post-translational modification, protein turnover, chaperones” (O), and “Cytoskeleton” (Z), were essentially underrepresented in the subset of extremely alkaline proteins (Fig. 4C). Also, a number of less abundant metabolic categories, such as “Cell cycle control and mitosis” (D), “Lipid transport and metabolism” (I), “Inorganic ion transport and metabolism” (P), etc., were underrepresented in this subset (Fig. 4A,B). In sum, the data obtained indicate that the extremely alkaline proteins of the human proteome are predominantly involved in RNA processing, translation, ribosomal organization and biogenesis.

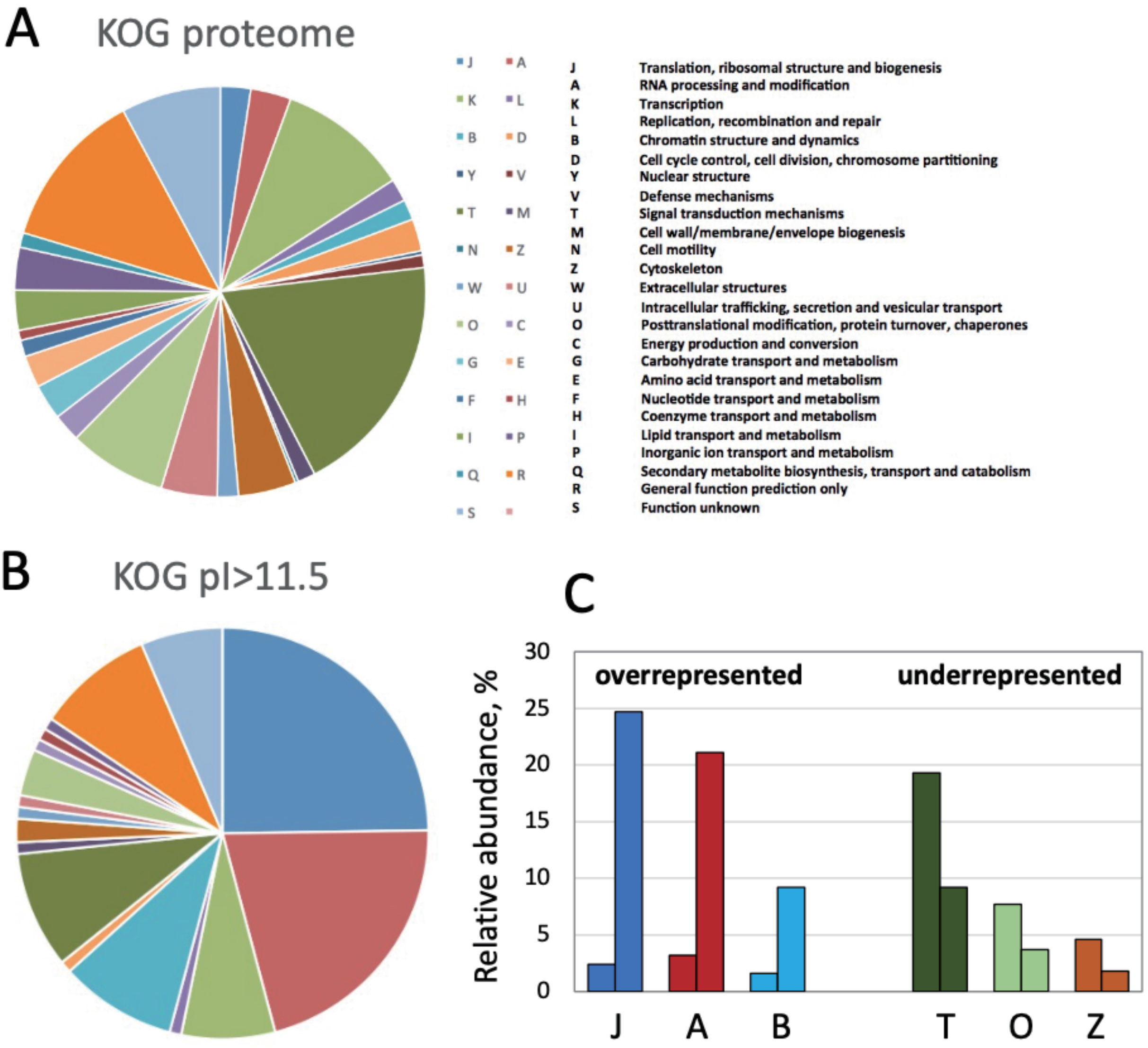

Fig 4. Assignments of a specific protein function for the studied subsets. Functional assignments were carried out for the complete dataset (**A**) and the subset of extremely alkaline proteins (**B**) of the human proteome. Specific functional categories of the individual proteins were assigned using the KOG classification algorithm. Relative abundance of the functional groups overrepresented or underrepresented in the subset of extremely alkaline proteins, as compared to the complete dataset, is shown in panel (**C**).

Nucleolar localization of extremely alkaline nuclear proteins

To get a deeper insight into the function of extremely alkaline proteins in the human proteome, the nucleolar localization signal (NoLS) was bioinformatically predicted in the amino acid sequences of these proteins. Remarkably, the proteins of the most abundant functional category “Information storage and processing” exhibited the highest probability of NoLS, exceeding 73%, whereas NoLS probability for the proteins in the categories of “Cellular processes and signaling” and “Metabolism” was essentially lower (Fig. 5A). The probability of NoLS for the whole subset of the extremely alkaline proteins was quite high too (57%), due to abundance of the functional category

“Information storage and processing” in this subset (Fig. 3B). Importantly, the proteins of nuclear localization in the extremely alkaline subset had the highest probability of NoLS (62%), whereas NoLS probability for mitochondrial proteins was low (23%) (Fig. 5B). The high probability of NoLS for the whole subset of the extremely alkaline proteins (57%), can be explained by the high abundance of nuclear proteins in this dataset (Fig. 2B). Altogether, these results suggest that the majority of the extremely alkaline nuclear proteins in the human proteome have nucleolar localization. This localization is consistent with the predominant functional assignments, such as “Translation, ribosomal structure and biogenesis” (J), “RNA processing and modification” (A), and “Chromatin structure and dynamic” (B), of the extremely alkaline proteins (Fig. 4B). Markedly, the nucleolus is known as the site of ribosomal biogenesis abundant with RNA and DNA.

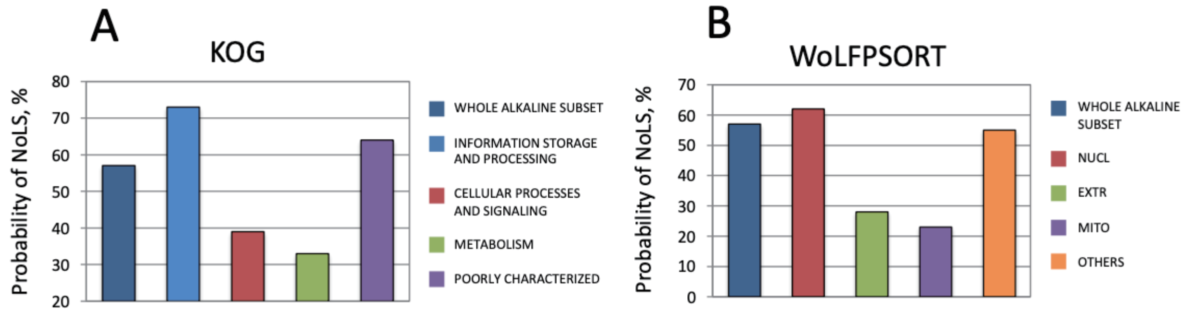

Fig 5. Prediction of the nucleolar localization signal, NoLS, in the subset of extremely alkaline proteins. Probabilities of NoLS in the proteins of different general function (KOG) and intracellular localization (WoLFPSORT) are presented in panels (**A**) and (**B**), respectively.

Conclusions

In the present study, calculative and predictive bioinformatics algorithms were used to designate pI values, intracellular locations, and functional categories of proteins in the human proteome. Specifically, a subset of extremely alkaline human proteins (pI>11.5), covering about 1.5% of the whole proteome, was scrutinized. Our study demonstrates that (i) the majority of the extremely alkaline proteins have nuclear or mitochondrial localization; (ii) the proteins of this subset are predominantly involved in RNA processing, ribosomal organization and biogenesis; (iii) most of the extremely alkaline nuclear proteins have nucleolar localization. A rationale behind the extreme alkalinity of the nucleolar proteins is presently unknown and requires investigation.

References

1. Schwartz, R., Ting, C. S., and King, J. (2001) Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. *Genome Res.* **11**, 703–709.
2. Knight, C. G., Kassen, R., Hebestreit. H., and Rainey, P. B. (2004) Global analysis of predicted proteomes: functional adaptation of physical properties. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 8390-8395.
3. Wu, S., Wan, P., Li, J., Li, D., Zhu, Y., and He, F. (2006) Multi-modality of pI distribution in whole proteome. *Proteomics* **6**, 449-455.
4. Carugo, O. (2007) Isoelectric points of multi-domain proteins. *Bioinformation* **2**, 101-104.
5. Kurotani, A., Tokmakov, A. A., Sato, K. I., Stefanov, V. E., Yamada, Y., and Sakurai, T. (2019). Localization-specific distributions of protein pI in human proteome are governed by local pH and membrane charge. *BMC Mol. Cell. Biol.*, *20*(1), 36.
6. Tokmakov, A., and Kurotani, A. (2021) Protein pI and intrcellular localization. *Front. Mol. Biosci.* 8:775736.
7. Ho, E., Hayen, A., and Wilkins, M. R. (2006). Characterisation of organellar proteomes: a guide to subcellular proteomic fractionation and analysis. *Proteomics*, *6*(21), 5746–5757.
8. Brett, C. L., Donowitz, M., and Rao, R. (2006) Does the proteome encode organellar pH? *FEBS Lett.* **580**, 717-719.

9. Kiraga, J., Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Biecek, P., Polak, N., Smolarczyk, K., et al. (2007) The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. *BMC Genomics* 8:163.

10. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152.

11. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R, Appel, R. D., and Bairoch, A. (2005) *Protein Identification and Analysis Tools on the ExPASy Server. The Proteomics Protocols Handbook*, Walker, J. M. (ed), pp. 571-607, Humana Press.

12. Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., and Nakai, K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* **35** (Web Server issue), W585-587.

13. Koonin, E. V, Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Krylov, D. M., Makarova, K. S., Mazumder, R., et al. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5**, R7.

14. Scott, M. S., Troshin, P. V., and Barton, G. J. (2011) NoD: a Nucleolar localization sequence detector for eukaryotic and viral proteins. *BMC Bioinformatics* 12:317.