

A Gene-centric Human Proteome Project

HUPO-THE HUMAN PROTEOME ORGANIZATION

The success of the Human Genome Project (1, 2) has provided a blueprint for the gene-encoded proteins potentially active in all of the hundreds of cell types that make up the human body. Yet we still have limited knowledge regarding a majority of the approximately 20,000 protein-coding human genes discovered through the genome project (3, 4). At present, about 8,000 (38%) of these genes lack experimental evidence at the protein level (UniProt), and for many others there is very little information related to protein abundance, distribution, subcellular localization, and function.

The diagnostic, prognostic, and therapeutic value of understanding human biology at the protein level argues for a systematic effort to map the human proteome. The proteomic space generated from these gene products is enormous, including up to a million different protein molecules derived by combinatorial recombination of DNA (immunoglobulins and T-cell receptors), alternative splicing of RNAs, and numerous protein modifications of various types that vary with time and with physiological, pathological, and pharmacological perturbations. Hochstrasser (5) therefore recently argued for a protein-centric human proteome project, driven by mass spectrometry technology focusing on the protein perturbations caused by human diseases. Our goal is to define clear endpoints of a Human Proteome Project, combining the strengths of complementary technology platforms.

We therefore propose a gene-centric approach to generate a human proteome map with an "information backbone" about the proteins expressed from each gene locus and to make this information publicly available with no restrictions, as was done with the genome sequence data, thereby facilitating in-depth studies to understand human biology and diseases. With further analogy with the genome project, the gene-centric human proteome map can be complemented with in-depth studies on protein variability with relevance to life stages and various diseases.

Reasonable end points of such a Human Proteome Project would be feasible within a limited time period and achievable without major paradigm shifts in technology. Taking into account recent major advances in mass spectrometry (6) and immunobased methods (7), we propose a systematic three-part approach to ensure that, for each predicted protein-coding gene, at least one of its major representative proteins will be characterized in the context of its major anatomical sites of expression, its abundance, and its interacting protein partners:

1. Protein parts list: the identification and characterization of at least one representative protein from every human

gene with its abundance and major modifications. This would define the backbone of a human proteome encyclopedia.

2. Protein distribution atlas: determination of protein profiles of at least one representative protein from every human gene in all major normal tissues and organs with single cell resolution as has been initiated (7). This effort already presents immunohistochemical characterization for 6,800 proteins corresponding to approximately one-third of the protein-encoded genes. The atlas also should include a subcellular localization map in which the relative distribution of the protein in various organelles and other subcellular structures is determined in selected human cell lines.
3. Protein pathway and network map: initial characterization of the transient and stable interactions and complexes of human proteins with other proteins (8) that contribute to cellular protein pathways. Networks of interacting proteins and pathways will be essential to describe biological functions at the molecular level, to understand disease processes, and to generate therapeutic targets; over time the interactions could be expanded to nucleic acids, lipids, and other molecules.

The protein profiling should ideally be performed using complementary mass spectrometry and protein capture technology platforms with proper standardization (9, 10) to allow comparative studies and with emphasis on quantification to enable systems biology analyses. In addition, knowledge from model organisms can be used to complement studies of humans as has been so productive for the genome. Complementary technology platforms, such as mRNA profiling, gene knock-outs, short interfering RNA silencing, green fluorescent protein fusions, and gene tagging, can be used with bioinformatics to validate and integrate the results. At least in the first phase, the various efforts can be pursued in a federated, decentralized manner involving many leading research groups in different regions of the world with coordination to avoid redundancy and to ensure standardization and completeness. In this way, each region can direct its funding to projects of relevance for its interests and needs. An important task will be to integrate the enormous data flow from these analyses; implementation of standard data exchange formats is essential (11).

In summary, we propose a coordinated, international effort to map the protein complement of the human genome. The effort will deliver a publicly available resource of protein profiling data, protein-specific reagents, and cDNA clones covering all the protein-coding genes of the human genome. With the aid of current and emerging advances in a well defined set of com-

plementary technology platforms, we envision a genome-wide human protein resource comprising the proteome parts list, the protein distribution atlas, and maps of protein molecular pathways, interactions, and networks. This approach provides a foundation for disease-oriented protein-centric studies by integrating information from gene and protein function with clearly defined end points. The output will be a major step to define the proteomic landscape in the human body, support the discovery of new diagnostic and therapeutic tools, and stimulate biological and medical research.

Call for inputs: HUPO has organized a small international working group to develop a detailed plan for the vision, goals, deliverables, timetable, launch, and funding of this Human Proteome Project. Open discussions are planned for regional HUPO meetings around the world, leading up to a roll-out at the HUPO World Congress in Sydney 19–23 September, 2010. Comments and questions from the proteomics community are welcome through a Contact us link at the HUPO website, www.hupo.org.

REFERENCES

- Lander, E. S., Linton, L. M., Birren, B., Nussbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczy, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissole, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickinson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramsier, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and Szustakowski, J. (2001) Initial sequence analysis of the human genome. *Nature* **409**, 860–921
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferreira, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, C., Kagan, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001) The sequence of the human genome. *Science* **291**, 1304–1351
- Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., and Lander, E. S. (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19428–19433
- UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.* **37**, D169–D174
- Hochstrasser, D. J. (2008) Should the human proteome project be gene- or protein-centric? *J. Proteome Res.* **7**, 5071
- Cox, J., and Mann, M. (2007) Is proteomics the new genomics. *Cell* **130**, 395–398
- Berglund, L., Björling, E., Oksvold, P., Fagerberg, L., Asplund, A., Szilyarto, C. A., Persson, A., Ottosson, J., Wernérus, H., Nilsson, P., Lundberg, E., Sivertsson, A., Navani, S., Wester, K., Kampf, C., Hober, S., Pontén, F., and Uhlén, M. (2008) A gene-centric human protein atlas for expression profiles based on antibodies. *Mol. Cell. Proteomics* **7**, 2019–2027
- Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albalá, J. S.,

- Lim, J., Fraughton, C., Llamasas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178
9. Hermjakob, H., and Apweiler, R. (2006) The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. *Expert Rev. Proteomics* **3**, 1–3
10. Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stümpflen, V., Ceol, A., Chatr-aryamontri, A., Armstrong, J., Wool-lard, P., Salama, J. J., Moore, S., Wojcik, J., Bader, G. D., Vidal, M., Cusick, M. E., Gerstein, M., Gavin, A. C., Superti-Furga, G., Greenblatt, J., Bader, J., Uetz, P., Tyers, M., Legrain, P., Fields, S., Mulder, N., Gilson, M., Niepmann, M., Burgoon, L., De Las Rivas, J., Prieto, C., Perreau, V. M., Hogue, C., Mewes, H. W., Apweiler, R., Xenarios, I., Eisenberg, D., Cesareni, G., and Hermjakob, H. (2007) The minimum information required for reporting a molecular interaction experiment. *Nat. Biotechnol.* **25**, 894–898
11. Omenn, G. S., Aebersold, R., and Paik, Y. K. (2009) The human plasma proteome project. *Proteomics* **9**, 4–6