D762–D766 *Nucleic Acids Research, 2009, Vol. 37, Database issue*
doi:10.1093/nar/gkn872

Human Gene and Protein Database (HGPD): a novel database presenting a large quantity of experiment-based results in human proteomics

Yukio Maruyama[1], Ai Wakamatsu[2,3,4], Yoshifumi Kawamura[1,2], Kouichi Kimura[5], Jun-ichi Yamamoto[3], Tetsuo Nishikawa[3,5,6], Yasutomo Kisu[2], Sumio Sugano[7], Naoki Goshima[2], Takao Isogai[3,4] and Nobuo Nomura[2,*]

[1]Japan Biological Informatics Consortium (JBIC), Aomi, Koto-ku, Tokyo 135-8073, [2]National Institute of Advanced Industrial Science and Technology (AIST), Aomi, Koto-ku, Tokyo 135-0064, [3]Reverse Proteomics Research Institute, Kisarazu, Chiba 292-0818, Japan, [4]Graduate School of Pharmaceutical Sciences, The University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, [5]Life Science Research Laboratory, Central Research Laboratory, Hitachi, Ltd., Kokubunji, Tokyo 185-8601, [6]Database Center for Life Science Research Organization of Information and Systems, Yayoi, Bunkyo-ku, Tokyo 113-0032 and [7]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Shiroganedai, Minato-ku, Tokyo 108-8639, Japan

Received August 14, 2008; Revised September 29, 2008; Accepted October 17, 2008

ABSTRACT

Completion of human genome sequencing has greatly accelerated functional genomic research. Full-length cDNA clones are essential experimental tools for functional analysis of human genes. In one of the projects of the New Energy and Industrial Technology Development Organization (NEDO) in Japan, the full-length human cDNA sequencing project (FLJ project), nucleotide sequences of approximately 30 000 human cDNA clones have been analyzed. The Gateway system is a versatile framework to construct a variety of expression clones for various experiments. We have constructed 33 275 human Gateway entry clones from full-length cDNAs, representing to our knowledge the largest collection in the world. Utilizing these clones with a highly efficient cell-free protein synthesis system based on wheat germ extract, we have systematically and comprehensively produced and analyzed human proteins *in vitro*. Sequence information for both amino acids and nucleotides of open reading frames of cDNAs cloned into Gateway entry clones and *in vitro* expression data using those clones can be retrieved from the Human Gene and Protein Database (HGPD, http://www.HGPD.jp). HGPD is a unique database that stores the information of a set of human Gateway entry clones and protein expression data and helps the user to search the Gateway entry clones.

INTRODUCTION

In 2003, complete sequences of the human genome were decoded by the human genome sequencing project (1). In postgenomic research, one of the most essential subjects is the functional and structural analysis of gene products (proteins). As access to full-length cDNA clones is critical for such work, many projects, such as the FLJ project (2,3), the Kazusa cDNA project (4), the US Mammalian Gene Collection (MGC) program (5), German (6), Chinese (7) and other cDNA projects have been executed to isolate as many full-length cDNAs with as high quality as possible. For comprehensive and high-throughput expression of human proteins, both full-length cDNA clones and a versatile system for using these clones are essential. For functional analysis of proteins, one often needs to fuse various tags at either the N- or C-termini, to adjust the reading frames of the open reading frame (ORF) and tags or to locate adequate regulatory sequences [promoters, enhancers, internal ribosomal entry sites (IRESes), etc.] close to the ORF. These manipulations can be extremely difficult when a large number of clones are being handled. The Gateway cloning system (Invitrogen, CA, USA) is based on versatile expression

*To whom correspondence should be addressed. Tel: +81 3 3599 8137; Fax: +81 3 3599 8141; Email: nomura.88@aist.go.jp
Correspondence may also be addressed to Takao Isogai, Tel: +81 3 5841 4775; Fax: +81 3 5841 4775; Email: tisogai@mol.f.u-tokyo.ac.jp; Naoki Goshima, Tel: +81 3 3599 8136; Fax: +81 3 3599 8141; Email: n-goshima@aist.go.jp

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

vectors and has the potential to overcome these barriers (8). We have therefore adopted Gateway technology and constructed 33 275 human Gateway entry clones that will serve as key resources for this versatile system. Sequence information of Gateway entry clones can be retrieved from the Human Gene and Protein Database (HGPD, http://www.HGPD.jp or http://HGPD.lifesciencedb.jp/) (Figure 1). ORFDB (http://orf.invitrogen.com/) (9) and the ORFeome collaboration (http://www.orfeomecollaboration.org/html) have been published as similar resources. Entry clones in ORFDB are only N-types which have a stop codon at the end of the ORF and are primarily dedicated for native or N-terminal fusion proteins, although one could produce native or C-terminal-fused protein with suppression technology (10). Lamesch *et al.* (11) have reported on the construction of 12 212 entry clones with which the ORFeome collaboration was formed. Since a large fraction of ORFeome clones are F-types that delete the stop codon for C-terminal fusion proteins, proteins that possess a functional domain at the C-terminus might not have full biological activity when expressed based on these clones. Therefore, one might need both N- and F-type entry clones. In our collection, both types have been prepared for 11 774 cDNAs, which means that our collection may have more flexibility for various *in vitro* and *in vivo* experiments.

Utilizing these clones with a highly efficient cell-free protein synthesis system featuring wheat germ (12), we have produced and analyzed 13 364 human proteins *in vitro*. The expression data can be retrieved from HGPD (Figure 1). HGPD manages and stores primary protein expression data, which differs from other databases, such as the Human Protein Reference Database (HPRD, http://www.hprd.org/) (13), Gene Ontology database (GO, http://www.geneontology.org/) (14), Universal Protein Resource (UniProt, http://www.pir.uniprot.org/) (15) or NCBI Entrez Gene (http://www.ncbi.nlm.nih.gov/) (16).

DATABASE CONTENTS

In HGPD, biological data such as *in vitro* expression data of human proteins are presented on the frame of cDNA clusters. To build the basic frame, sequences of FLJ cDNAs and others deposited in public databases (Human ESTs, RefSeq, Ensembl, MGC, etc.) are assembled onto the genome sequences (Table 1). Information for human Gateway entry clones is presented with the source cDNAs. The specific features of the HGPD that we would like to emphasize are that it contains (i) the world largest collection of Gateway entry clones, (ii) arrangement of both N- and F-type entry clones, and especially, (iii) SDS–PAGE patterns of proteins expressed in the cell-free wheat germ extract system ('PE' shown in Figure 1).

Gateway entry clones

To facilitate utilization of full-length cDNA clones, we have adopted the versatile Gateway expression system which offers high-throughput gene transfer technology for functional gene analysis and protein expression. For conversion to entry clones, we selected an ORF region in each cDNA meeting one of the following criteria: (i) ORFs-encoding products $\geq$150 aa [although the longest ORF starting with an AUG codon has highest priority, the selected ORF is finally determined by taking into consideration homology search results of shorter ORFs with BLASTX(nr) and BLASTP against SwissProt and RefSeq databases); (ii) both '149 aa $\geq$ORF $\geq$100 aa' and 'ORF with an ATGpr value (17) $\geq$0.4'; (iii) both '100 aa >ORF' and 'known gene'. Those ORF regions were PCR amplified with attB sequences of the Gateway system at both ends. Then those fragments were recombined with attP sequences of the Gateway donor vector pDONR201 (Invitrogen). Eventually, we constructed 33 275 Gateway entry clones utilizing FLJ clones as major cDNA resources. Sequence information, such as amino acid and nucleotide sequences of ORF regions and sequence differences of Gateway entry clones from source cDNAs are presented in the 'GW: Gateway Summary' page (for help, see http://hgpd.hinv.jp/sys_info/help.html#w120_gw). The details for construction and usage of entry clones will be published elsewhere (18).

Gateway entry clones are available from NITE Biological Resource Center (NBRC), Department of Biotechnology, National Institute of Technology and Evaluation. Distribution of clones by NBRC requires the signing of an MTA by both private companies and academic institutions. Distribution charges will be 30 000 and 15 000 Yen (JPN; approx. US$300 and $150, respectively) per clone for private companies and academic institutions, respectively. More information is available through the 'clone inquiry' page (http://hgpd.hinv.jp/sys_info/order_clone.html) of HGPD or the notice page (http://www.nbrc.nite.go.jp/e/hgentry-e.html) of NBRC.

SDS–PAGE patterns of human proteins synthesized *in vitro*

The Gateway system is a versatile expression vector system that is adequate for handling large numbers of clones. For expression of large numbers of human proteins, we adopted the wheat germ cell-free protein synthesis system. In addition, we devised a new procedure to prepare template DNA for transcription, which makes the step simpler and more efficient. By applying those protocols, we expressed 13 364 human proteins with a C-terminal V5 or His tag and analyzed them using SDS–PAGE. Expression patterns of proteins in both the total and supernatant fractions are displayed in the 'PE: Protein Expression' page (for details, see http://hgpd.hinv.jp/sys_info/help.html#w120_pe). Essentially all of the human proteins analyzed in our work were shown to be expressed. This implies that *in vitro* cell-free systems using wheat germ extract offer a very efficient system for protein production.

Computational analysis of individual cDNA sequences with BLAST, Pfam, PROSITE, PSORT, SignalP, SOSUI and GO

Functional motifs and domains, subcellular localization information, leader sequences and transmembrane

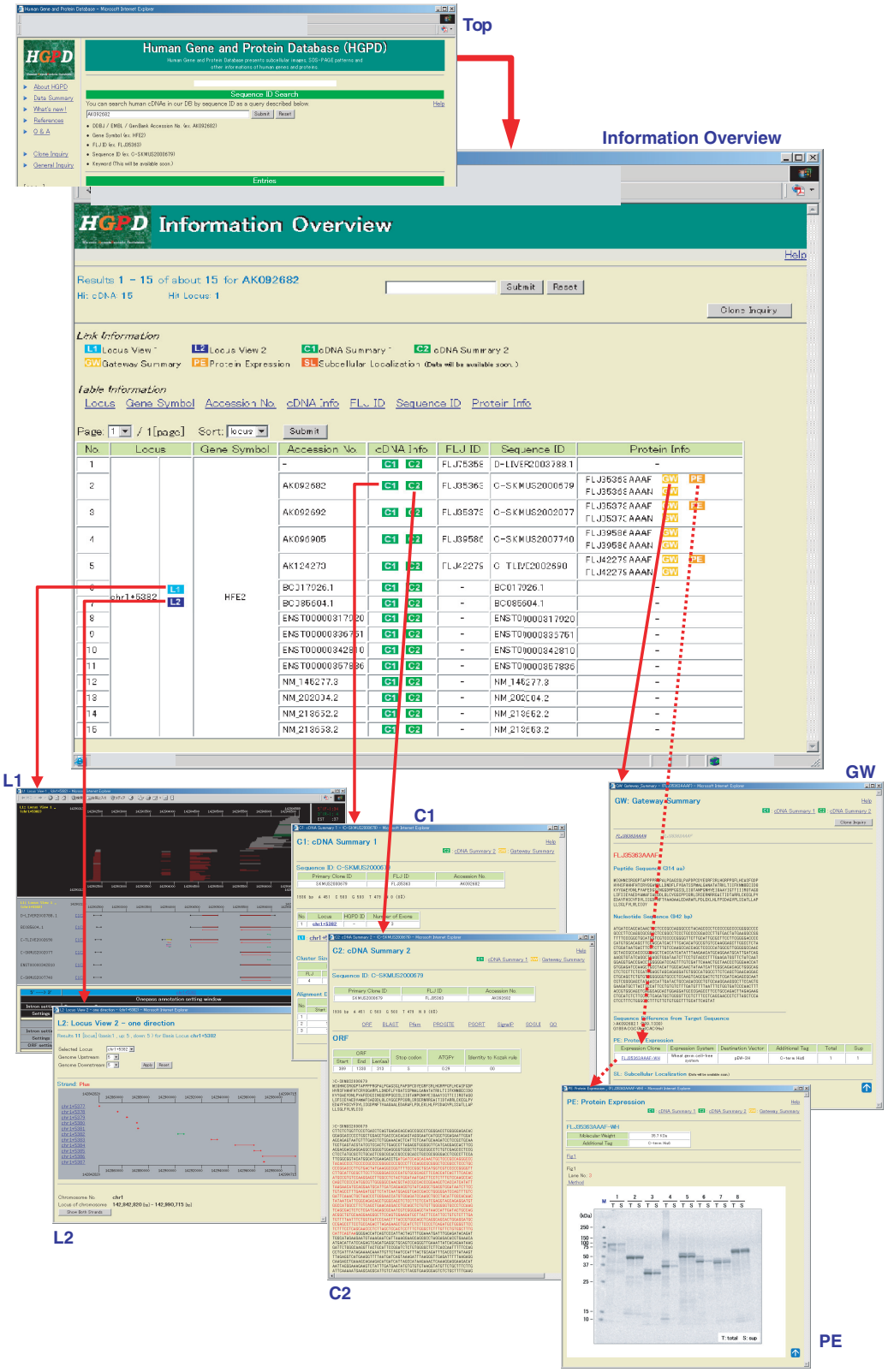

Figure 1. Search flow in HGPD. Each representative page of HGPD is shown: Top, top page. After entering a proper ID, such as 'DDBJ/EMBL/GenBank Accession No.', 'Ensembl Transcript No.', 'Gene Symbol', 'FLJ ID' and 'Sequence ID', the 'Information Overview' window will emerge. It presents a summary of all information on the cluster to which the queried cDNA clone belongs. Search results for 'AK092682' (DDBJ/EMBL/GenBank Accession No., AK092682; FLJ ID, FLJ35363; Sequence ID, C-SKMUS2000679) are presented as an example. A 'PE' button opens a 'Protein Expression' window through a GW page, which is indicated by dotted arrows. An L2′ window which is linked with an L2 window and can be used for 'search by chromosome coordinates' is not shown (for details, see http://hgpd.hinv.jp/sys_info/help.html#l2b).

Table 1. Entries of HGPD

Dataset	Number of data
Gateway Entry Clones _ N-type[a]	12 754
Gateway Entry Clones _ F-type[b]	20 521
In vitro Protein Expression (SDS–PAGE) Patterns[c]	13 364
FLJ cDNAs[d]	35 083
Public Database cDNAs (including RefSeq, Ensembl, DKFZ and others)	112 992
FLJ_ESTs[e]	1 430 438
Public_ESTs	3 862 807

[a]Entry clones with a naturally occuring stop codon.
[b]Entry clones without a stop codon for adding a tag at the C-terminal end.
[c]As the number of FLJ cDNAs.
[d]Number of published clones: 30 063, unpublished clones: 5020.
[e]Deposited by the FLJ project.

domains were inferred using BLAST, Pfam (http://www.sanger.ac.uk/Software/Pfam/), PROSITE (http://www.expasy.ch/prosite/), PSORT (http://psort.hgc.jp/), SignalP (http://www.cbs.dtu.dk/services/SignalP/), SOSUI (http://bp.nuap.nagoya-u.ac.jp/sosui/sosuimenu0.html) and GO.

Mapping and clustering of cDNA clones

Local alignments between human cDNAs and human genome sequences (UCSC hg17 NCBI Build 35) were calculated using megablast (http://www.ncbi.nih.gov/blast). Initially, the alignment with the highest score was selected and a single locus was assigned for each cDNA. Those cDNAs with sequences overlapping not less than 1 base at the same locus and strand were defined as constituting the same cluster. All entries cataloged in HGPD are presented in Table 1.

WEB INTERFACE

The search flow of HGPD is illustrated in Figure 1. The top page (http://hgpd.hinv.jp/sys_info/help.html#id_search) of the HGPD viewer is represented in the upmost part of Figure 1. To begin the search, the ID number (in a definitive or degenerated form) such as DDBJ/EMBL/GenBank accession number, Ensembl transcript number, Gene Symbol, FLJ ID or Sequence ID is entered into the text box. When a query hits the data in HGPD, an 'Information Overview' page comes out. It shows all data concerning all members clustered with a queried sequence. In addition, all information stored in HGPD for searched clusters and cDNA clones is documented on the page. The 'Locus' column represents the cluster ID obtained by genome mapping of all the cDNA sequences, including expressed sequence tags. Buttons 'L1' and 'L2' are linked with 'L1: Locus View 1' (http://hgpd.hinv.jp/sys_info/help.html#w015) and 'L2: Locus View 2' (http://hgpd.hinv.jp/sys_info/help.html#w022), respectively. The 'Gene Symbol' column represents the official symbol appearing in the Entrez Gene database for each cDNA clone. cDNA clones that have not been assigned a 'Gene Symbol' are designated as '-'. The 'Accession No.' column represents the registered ID in the public database for each cDNA clone. Buttons 'C1' and 'C2' in the 'cDNA Info' column are linked to 'C1: cDNA Summary 1' and 'C2: cDNA Summary 2' (for details, see http://hgpd.hinv.jp/sys_info/help.html#w013 and http://hgpd.hinv.jp/sys_info/help.html#w014 for C1 and C2, respectively). Information on cDNA clones, including sequences and homology search results, is presented on the 'cDNA Summary 1' and 'cDNA Summary 2' pages. The 'FLJ ID' column indicates the FLJ ID number of the FLJ cDNA clone. Any cDNA clone that has not been assigned an FLJ number is designated as '-'. FLJ clones were eventually found to have three kinds of IDs: 'DDBJ/EMBL/GenBank Accession No.', 'FLJ ID' and 'Primary Clone ID'. The 'Sequence ID' column shows the ID of a sequence of a cDNA clone. For sequences of cDNAs other than FLJ cDNAs, an accession number for DDBJ/EMBL/GenBank is depicted. The column 'Protein Info' is linked to information on expressed proteins using Gateway entry clones. A 'GW' button is linked with sequence information on entry clones and a 'PE' button is linked with protein expression through a 'GW: Gateway Summary' page.

In the search flow of HGPD, some links open new windows and other links load in the current window (http://hgpd.hinv.jp/sys_info/help.html#search_flow). Windows that show various data (C1, C2, GW and PE) focusing on a single cDNA clone open in the current window, as translocation can be essentially reversible (one versus one). Other windows which display multiple clones or clusters ('Information Overview', L1, L2 and L2′) will in principle open a new window when transferred, as translocation is usually irreversible (one versus multiple).

Data for amino acid and nucleotide sequences of ORFs cloned into Gateway entry clones, summary of protein expression and others can be downloaded from the top page of HGPD (http://hgpd.hinv.jp/sys_info/download.html).

FUTURE DEVELOPMENTS

Several modifications in browser interface will be done. (i) The database will be updated by next spring to correspond to UCSC hg18/NCBI build 36. (ii) Various search interfaces will be introduced in a future version.

Information on about 18 000 more human entry clones will be included shortly, which will put the cumulative number of our collection at 50 000. Fourteen thousand entries on protein expression data in *Escherichia coli* will also be presented in HGPD. Additionally, data for subcellular localization for 14 000 expressed human proteins, which have been examined in HeLa cells, are being processed for publication.

ACKNOWLEDGEMENTS

We thank the Helix Research Institute and the Research Association for Biotechnology for FLJ cDNA clones.

FUNDING

New Energy and Industrial Technology Development Organization 'Functional Analysis of Human Proteins and its Application' project and intramural research grants of National Institute of Advanced Industrial Science and Technology. Funding for open access charge: Japan Biological Informatics Consortium.

Conflict of interest statement. None declared.

REFERENCES

1. International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
2. Ota,T., Suzuki,Y., Nishikawa,T., Otsuki,T., Sugiyama,T., Irie,R., Wakamatsu,A., Hayashi,K., Sato,H., Nagai,K. *et al*. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet*., **36**, 40–45.
3. Kimura,K., Wakamatsu,A., Suzuki,Y., Ota,T., Nishikawa,T., Yamashita,R., Yamamoto,J., Sekine,M., Tsuritani,K., Wakaguri,H. *et al*. (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res*., **16**, 55–65.
4. Nomura,N., Miyajima,N., Sazuka,T., Tanaka,A., Kawarabayasi,Y., Sato,S., Nagase,T., Seki,N., Ishikawa,K. and Tabata,S. (1994) Prediction of the coding sequences of unidentified human genes. I. The coding sequences of 40 new genes (KIAA0001-KIAA0040) deduced by analysis of randomly sampled cdna clones from human immature myeloid cell line KG-1. *DNA Res*., **1**, 27–35.
5. Mammalian Gene Collection Program Team (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA*, **99**, 16899–16903.
6. Wiemann,S., Weil,B., Wellenreuther,R., Gassenhuber,J., Glassl,S., Ansorge,W., Böcher,M., Blöcker,H., Bauersachs,S., Blum,H. *et al*. (2001) Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res*., **11**, 422–435.
7. Hu,R., Han,Z., Song,H., Peng,Y., Huang,Q., Ren,S., Gu,Y., Huang,C., Li,Y., Jiang,C. *et al*. (2000) Gene expression profiling in the human hypothalamus-pituitary-adrenal axis and full-length cDNA cloning. *Proc. Natl Acad. Sci. USA*, **97**, 9543–9548.
8. Hartley,J., Temple,G. and Brasch,M. (2000) DNA cloning using in vitro site-specific recombination. *Genome Res*., **10**, 1788–1795.
9. Liang,F., Matrubutham,U., Parvizi,B., Yen,J., Duan,D., Michandani,J., Hashima,S., Nguyen,U., Ubil,E, Loewenheim,J. *et al*. (2004) ORFDB: an information resource linking scientific content to a high-quality open reading frame (ORF) collection. *Nucleic Acids Res*., **32**, D595–D599.
10. Drabkin,H.J., Park,H.J. and RajBhandary,H.L. (1996) Amber suppression in mammalian cells dependent upon of an *Escherichia coli* aminoacyl-tRNA synthetase gene. *Mol. Cell Biol*., **16**, 907–913.
11. Lamesch,P., Li,N., Milstein,S., Fan,C., Hao,T., Szabo,G., Hu,Z., Venkatesan,K., Bethel,G., Martin,P. *et al*. (2007) hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. *Genomics*, **89**, 307–315.
12. Sawasaki,T., Ogasawara,T., Morishita,R. and Endo,Y (2002) A cell-free protein synthesis system for high-throughput proteomics. *Proc. Natl Acad. Sci. USA*, **99**, 14652–14657.
13. Peri,S., Navarro,J., Amanchy,R., Kristiansen,T., Jonnalagadda,C., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T., Gronborg,M. *et al*. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*., **13**, 2363–2371.
14. Ashburner,M., Ball,C., Blake,J., Botstein,D., Butler,H., Cherry,M., Davis,A., Dolinski,K., Dwight,S., Eppig,J. *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.
15. The UniProt Consortium. (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **35**, D193–D197.
16. Maglott,D., Ostell,J., Pruitt,K. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **35**, D26–D31.
17. Nishikawa,T., Ota,T. and Isogai,T. (2000) Prediction whether a human cDNA sequence contains initiation codon by combining statistical information and similarity with protein sequences. *Bioinformatics*, **16**, 960–967.
18. Goshima,N., Kawamura,Y., Fukumoto,A., Miura,A., Honma,R., Satoh,R., Wakamatsu,A., Yamamoto,J.-i., Kimura,K., Nishikawa,T., *et al*. (2008) Human protein factory for converting the transcriptome into an in vitro–expressed proteome. *Nat. Methods*, **5**, 1011–1017.