BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 21 2007, pages 2934–2941
doi:10.1093/bioinformatics/btm430

Databases and ontologies

Construction of an open-access database that integrates cross-reference information from the transcriptome and proteome of immune cells

Atsushi Hijikata[1], Hiroshi Kitamura[1], Yayoi Kimura[1], Ryo Yokoyama[1], Yuichi Aiba[1], Yanyuan Bao[1], Shigeharu Fujita[1], Koji Hase[1], Shohei Hori[1], Yasuyuki Ishii[1], Osami Kanagawa[1], Hiroshi Kawamoto[1], Kazuya Kawano[1], Haruhiko Koseki[1], Masato Kubo[1], Ai Kurita-Miki[1], Tomohiro Kurosaki[1], Kyoko Masuda[1], Mitsumasa Nakata[1], Keisuke Oboki[1], Hiroshi Ohno[1], Mariko Okamoto[1], Yoshimichi Okayama[1], Jiyang O-Wang[1], Hirohisa Saito[1], Takashi Saito[1], Machie Sakuma[1], Katsuaki Sato[1], Kaori Sato[1], Ken-ichiro Seino[1], Ruka Setoguchi[1], Yuki Tamura[1], Masato Tanaka[1], Masaru Taniguchi[1], Ichiro Taniuchi[1], Annabelle Teng[1], Takeshi Watanabe[1], Hiroshi Watarai[1], Sho Yamasaki[1] and Osamu Ohara[1,2,*]

[1]RIKEN Research Center for Allergy and Immunology, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045 and [2]Department of Human Genome Technology, Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan

Received on March 7, 2007; revised on July 31, 2007; accepted on August 17, 2007

Advance Access publication September 24, 2007

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Although a huge amount of mammalian genomic data does become publicly available, there are still hurdles for biologists to overcome before such data can be fully exploited. One of the challenges for gaining biological insight from genomic data has been the inability to cross-reference transcriptomic and proteomic data using a single informational platform. To address this, we constructed an open-access database that enabled us to cross-reference transcriptomic and proteomic data obtained from immune cells.

Results: The database, named RefDIC (Reference genomics Database of Immune Cells), currently contains: (i) quantitative mRNA profiles for human and mouse immune cells/tissues obtained using Affymetrix GeneChip technology; (ii) quantitative protein profiles for mouse immune cells obtained using two-dimensional gel electrophoresis (2-DE) followed by image analysis and mass spectrometry and (iii) various visualization tools to cross-reference the mRNA and protein profiles of immune cells. RefDIC is the first open-access database for immunogenomics and serves as an important information-sharing platform, enabling a focused genomic approach in immunology.

Availability: All raw data and information can be accessed from http://refdic.rcai.riken.jp/. The microarray data is also available at http://cibex.nig.ac.jp/ under CIBEX accession no. CBX19, and http://www.ebi.ac.uk/pride/ under PRIDE accession numbers 2354–2378 and 2414.

Contact: hijikata@rcai.riken.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

*To whom correspondence should be addressed.

1 INTRODUCTION

When a complex biological event is to be explored, a data-driven approach is widely accepted as a powerful alternative to a conventional hypothesis-driven one (Smalheiser, 2002). However, the data-driven approach cannot be achieved without high-quality genomic data. In this regard, DNA microarray technology has opened the way to the comprehensive collection of large amounts of data regarding genome-wide gene expression profiles, and is now recognized as a standard tool for the characterization of biological systems at the mRNA level. The description of biological systems with tens of thousands of different mRNA expression levels is highly sensitive to the state of those systems, and has enabled us to discover a number of mRNA biomarkers for various biological events (Abbas *et al.*, 2005; Komor *et al.*, 2005; Su *et al.*, 2004; Zheng *et al.*, 2006). However, the discovery of such biomarkers does not necessarily address the molecular mechanisms underlying the observed biological events. This is mainly because mRNA levels do not always have direct relevance to biological phenomena in general; mRNA is a template of protein synthesis and is not a functional element in itself in most cases. In this respect, protein profiles must have a greater relevance to biological events than mRNA profiles because proteins govern them directly. Thus, genome-wide characterization of the biological system at the protein level is widely considered to be the next goal to achieve, although this in itself is challenging due to technological limitations faced at present (Cutler, 2003). The low availability of quantitative proteomic data is one of bottlenecks when

analyzing a biological system from an 'omics' viewpoint. To address this problem, we have recently reported an approach for generating quantitative proteomic maps based on two-dimensional gel electrophoresis (2-DE) and have attempted to expand the protein profiling data (Kimura *et al*., 2006).

As more mRNA and protein profile data has accumulated, the demand for a cross-referencing database has increased significantly. We have therefore developed a web-based platform for sharing specialized immune cell mRNA and protein profile data. For this study, all of the profile data was newly obtained following well-controlled protocols. Affymetrix GeneChip DNA microarray technology was utilized for obtaining mRNA profiles and 2-DE for quantitative protein profiling. In order to integrate the 'omics' data obtained for various immune cells, we annotated each sample with controlled vocabularies and implemented a relational database that included the quantitative mRNA and protein profiling data. The Reference genomics Database of Immune Cells (RefDIC) offers a web-based query interface and user-friendly data visualizing facilities, and also allows all of the raw data to be downloaded. RefDIC could serve as a solid reference for the transcriptome and proteome of immune cells, and hence could greatly facilitate the identification of immunologically important genes and proteins that are involved in various immune responses, through cross-referencing quantitative mRNA and protein profiling data.

2 SYSTEMS AND METHODS

2.1 Samples of immune cells and tissues

In this study, we analyzed a number of samples for mRNA and/or protein profiling including those derived from various tissues and immune cells extracted from human and mouse biopsies, such as T lymphocytes, B lymphocytes, natural killer cells, natural killer T (NKT) cells, dendritic cells, macrophages, mast cells and intestinal epithelial cells, as well as several popular lymphoid or myeloid cell lines. Detailed information regarding these samples and their preparation are available in sample attribute tables, which are accessible through clicking each sample identifier, termed 'cellID', found in the data sets section of the website (http://refdic.rcai.riken.jp/dataset.cgi).

2.2 Extracting data from public databases

For the integration of mRNA and protein profiling data for immune cells, we took advantage of the Entrez Gene database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene) at the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov) because it provides solid references for various lines of genomic information, such as transcript and protein sequences, and Gene Ontology (GO) terms with unique, stable and traceable gene identifiers (Maglott *et al*., 2007). The sequence data for transcripts and proteins were extracted from RefSeq, GenBank and UniGene databases maintained by NCBI. Additionally, protein sequence data was extracted from the International Protein Index (IPI, http://www.ebi.ac.uk/IPI/IPIhelp.html) served by the European Bioinformatics Institute (EBI, http://www.ebi.ac.uk/), since IPI is designed for complete non-redundant data sets for human, mouse and rat proteomes (Kersey *et al*., 2004). The protein domain data was extracted from Pfam (http://www.sanger.ac.uk/Software/Pfam/). Because entries in the RefSeq protein database do not have external links to Pfam, we assigned Pfam domains in the respective RefSeq peptide entries using HMMER (Eddy, 1998), with the threshold *E*-value set to 0.1. All sequence data for probe sets on Affymetrix GeneChip expression arrays was downloaded from the Affymetrix website (http://www.affymetrix.com/index.affx). Human and mouse genome sequence data (Build 36) was also downloaded from the NCBI website. We collected publicly available microarray experimental data from Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress).

2.3 Microarray experiments

Total RNA was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) and/or an RNAeasy kit (Qiagen, Hilden, Germany). The RNA integrity was assessed using a bioanalyzer (Agilent Technologies Inc., Palo Alto, CA, USA). Samples whose RNA integrity number was greater than 7.0 were used for mRNA profiling by microarray analysis. cDNA synthesis, cRNA amplification, biotinylation and fragmentation were performed with a One-Cycle Target Labeling Kit (Affymetrix, Santa Clara, CA, USA). Twenty micrograms of labeled target RNA was hybridized with Mouse Genome 430 2.0 (Mouse430_2) or 430A 2.0 (Mouse430A_2), or Human Genome U133 Plus 2.0 (HG-U133_Plus2) GeneChip expression arrays (Affymetrix) at 45°C for 16 h, as described in the manufacturer's instructions. Washing stages and streptavidin-phycoerythrin staining were conducted using a GeneChip Fluidics Station (Affymetrix). Subsequently, the chips were scanned using a GeneChip Scanner 3000 (Affymetrix). Array data was normalized using either MAS5 (Hubbell *et al*., 2002) or gcRMA (Irizarry *et al*., 2003) algorithms. All of the microarray data were deposited to CIBEX with an accession number of CBX19 (http://cibex.nig.ac.jp/index.jsp).

2.4 Gene annotation of Affymetrix probe sets

Each probe set on the Affymetrix arrays consists of 22 oligonucleotides, each 25-bases long. Half of these probes were designed as a 'perfect match' (PM) for a specific transcript. To identify which transcript could be probed by each probe set, the 11 PM sequences were subjected to a Basic Local Alignment Search Tool (BLAST) search (Altschul *et al*., 1990) against RefSeq (39 179 human and 47 930 mouse cDNA sequences), GenBank (128 863 human and 147 850 mouse cDNA sequences) and UniGene [6 988 853 human and 4 277 970 mouse expressed sequence tags (EST) sequences] databases (updated on 1 November 2006). In this study, when the nucleotide sequences for more than 9 of the 11 probes in the set matched that of a given transcript perfectly, we considered that the probe set targeted this particular transcript. If they matched with a complementary sequence of a given transcript or with multiple transcript sequences originating from different gene loci, or if they failed to match any transcript in any of the databases, then it was concluded that the probe set targeted an antisense transcript of a known gene, or was a cross-hybridizing or non-informational probe set, respectively. Consequently, the same Entrez GeneID was given to a probe set targeting a specific transcribed sequence as for one targeting the corresponding transcript. Based on the results of the BLAST search, we classified all of the probe sets into six categories: categories A, B and C included the probe sets targeting specific RefSeq transcripts, specific GenBank transcripts (i.e. present in GenBank but not in RefSeq), and specific transcripts found only in the UniGene database, respectively; category D included those sets targeting antisense transcripts; category E consisted of cross-hybridizing probe sets and category X comprised non-informational probe sets.

2.5 Two-dimensional gel-based proteome experiments

Quantitative protein profiling by 2-DE followed by gel image analyses and mass spectrometry was performed essentially as described (Kimura *et al*., 2006) with a few modifications (Kimura *et al*., manuscript in preparation). Briefly, whole-cell protein samples (250 μg) were

separated by isoelectric focusing (IEF) in the first dimension and by SDS-PAGE in the second dimension, and subsequently the proteins in the gel were stained with SYPRO Ruby (Molecular Probes, Eugene, OR, USA). The 2-DE gel images were scanned using a ProXPRESS fluorescent imager (PerkinElmer, Inc., Boston, MA, USA) and analyzed using Progenesis Workstation (Nonlinear Dynamics Ltd, Newcastle upon Tyne, UK). The protein levels for each spot on a given gel were normalized by median centering. In order to identify the proteins in each spot on the gels, the mass spectra of digest fragments originating from the proteins excised from each gel were obtained by peptide mass fingerprinting (PMF) and/or MS/MS methods and were searched against the IPI database (Version 3.21; 51 432 mouse protein sequences) using the MASCOT Version 1.8.06 (Matrix Science, London, UK). To define a spot on a gel that corresponded to one on another gel, all gel images were superimposed. When the position of a spot on a gel matched that of one on another, we considered them provisionally as identical protein spots. Each protein was assigned to one of five classes (A–E) based on reliability of identification, primarily according to the degree of confidence in the database search results for PMF and/or MS/MS analysis data (Kimura *et al.*, 2006). The reliability for classes A (conclusive) and B (most likely) was based solely on the MS data. Protein identification in class C was supported by comigration of the previously identified proteins on 2-DE and was consistent with the predicted MS data for the identified proteins, although the MS data alone did not allow us to identify the protein with sufficient confidence. On the other hand, proteins identified in class D were supported only by their electrophoretic behaviors on 2-DE, while proteins in class E remained uncharacterized even after we combined 2-DE and MS data. All of the proteomic data were deposited into the PRIDE database with accession numbers of 2354–2378 and 2414 (http://www.ebi.ac.uk/pride/).

2.6 Correlation analysis of mRNA and protein profiles

For correlation analysis of mRNA and protein levels from the same gene in different samples, Pearson's correlation coefficients were used. Values for the signal intensity of the probe sets were taken as expression levels of transcripts from the corresponding genes. When gene products from a single gene produced multiple protein spots on a 2-DE gel, the sum of the volumes of these protein spots were taken as the volume of these gene products in this study. When a protein spot was found to include two or more proteins, it was excluded from the analysis. The significance of the correlation coefficients was tested using the one-tailed t-distribution with $(n-2)$ degrees of freedom.

2.7 Web database platform

A web server for the RefDIC is on a machine running CentOS (http://www.centos.org/) and the Apache web server (http://apache.org/). All scripts for data querying, retrieving and visualization were written in Perl. A MySQL 4.1x server (http://www.mysql.com/) is used as the storage engine for the database.

3 IMPLEMENTATION

3.1 Database structure and design

RefDIC segregates data into three components of the MySQL tables: '*Sample*', '*Microarray experiments*' and '*Proteome experiments*' (Supplementary Fig. S1). Each immune cell sample is associated internally with a unique identifier, named cell ID, for the determination of connecting lines of information regarding characteristic features of a given sample and the corresponding quantitative mRNA and protein profiling data. The tables of 'SampleSource', 'SampleTreatment' and 'TreatmentDescription' in the '*Sample*' component provide records of information regarding the characteristic features of a given sample, using controlled vocabulary to ensure that samples can be distinguished. The '*Microarray experiments*' component has three tables: 'MicroarrayExperiment', 'MAS5_data' and 'gcRMA_data'. The 'MicroarrayExperiment' table is a record of information regarding the microarray experimental procedures. The 'MAS5_data' and 'gcRMA_data' tables are records of the values for the signal intensity of the probe sets on Affymetrix GeneChip arrays. The '*Proteome experiments*' component has four tables: '2DEgelExperiment', '2DEgel_data', 'protein_identification' and 'spot_matching'. The '2DEgelExperiment' table is a record of the number of protein spots that could be quantified on the 2-DE gel. The '2DEgel_data' table is a record of the information regarding respective protein spots, such as the normalized protein level, isoelectric point, molecular weight, etc. The 'protein_identification' table is a record of the information regarding identified proteins for the respective protein spots and the 'spot_matching' table is a record of the relationship between corresponding protein spots, identified using the procedure described in the Systems and Methods section. The mRNA and protein profiling data for respective genes from the microarray and 2-DE-based proteomic experiments could be linked together with Entrez GeneIDs as a central hub on the MySQL database.

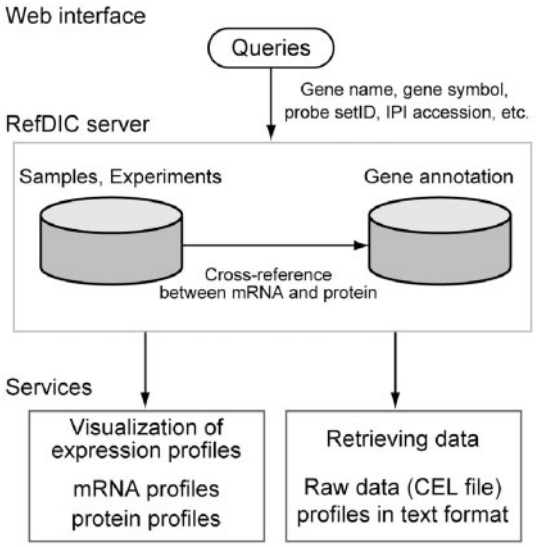

Fig. 1. An overview of the web-based query interface and retrieval system for the RefDIC server. A user can submit various types of query to view gene expression profiles and to retrieve the experimental raw data, such as CEL files, from the web interface.

3.2 Using a web-based query interface to access quantitative data

Figure 1 gives an overview of the web-based query interface and retrieval system and how RefDIC responds when a user submits a query. RefDIC provides mRNA and protein

profiling data for the same sample. This data is cross-referenced on the relational database and thus the web interface allows users to visualize both the mRNA and protein profiling data. RefDIC also provides the experimental raw data regarding CEL files for the microarray and mass spectra for the proteomic analyses, respectively. The raw data is accessible in 'Data downloads' on the RefDIC website.

4 RESULTS AND DISCUSSION

4.1 Data statistics for RefDIC

RefDIC is comprised of quantitative transcriptomic and proteomic data for immune cells. For transcriptomic data, a total of 125 and 34 microarray data from 60 to 21 different subsets of immune cells and tissues from mouse and human, respectively (Table 1), have been newly obtained and stored in the database. Table 2 summarizes the results for the annotation and classification of Affymetrix GeneChip probe sets. Based on our annotation system we found that 39 406 and 43 835 probe sets on Mouse430_2 and HG-U133_Plus2, respectively, could detect specific transcripts from 20 623 and 19 300 distinct mouse and human genes, respectively. For proteomic data, 23 2-DE gel images from 21 different subsets of immune cells and 2 subsets of epithelial cells were obtained and stored for the mouse (Table 1). We could identify at least 963 protein spots on each gel image, and 435 proteins in total from distinct genes (Table 3 and Supplementary Table S1). We quantified the amount of protein in these spots by gel image analysis. However, because most of our quantitative proteome data were obtained by a single 2-DE experimental run, the quantitative protein data must be interpreted with caution (Fievet *et al*., 2004; Gustafsson *et al*., 2004; Mahon *et al*., 2001). According to our previous study, the average coefficient of variation for the protein spot quantitative values obtained using essentially same protocol as in this study was 26.8% (range 3.6–105%; Kimura *et al*., 2006). For convenience, the actual gel images used for quantification are accessible by clicking on matchIDs found in the proteome profile section. The current availability of mRNA and protein profile data can be checked in 'Statistics' in the data set section of RefDIC (http://refdic.rcai.riken.jp/dataset.cgi).

4.2 Evaluation of the probe set annotation for Affymetrix GeneChip arrays and their relevance to protein profiling data

As some groups have already pointed out, there are problems associated with Affymetrix GeneChip probe set annotation (Dai *et al*., 2005; Harbig *et al*., 2005; Zhang, *et al*., 2005). Because Affymetrix utilized information that was incomplete at the time of GeneChip design, the current GeneChip probe set design could include some discrepancies when compared to the most up to date genomic data. Furthermore, although the

Table 1. The data sets represented in RefDIC

Cell type	Mouse			Human	
	Number of subsets	Number of microarrays	Number of 2-DE	Number of subsets	Number of microarrays
B lymphocyte	15	32	13	1	4
T lymphocyte	14	31	4	7	13
Dendritic cell	6	17	2	2	2
Macrophage	4	6(6)	2	–	–
Natural killer cell	1	2	–	2	4
NKT cell	2	3	–	–	–
Mast cell	1	6	–	–	–
Monocyte	–	–	–	1	2
Leukocyte (mixture)	1	2	–	–	–
Epithelial cell	3	5	2	–	–
Pancreatic beta cell	1	1	–	1	1
Spleen	1	1	–	1	1
Thymus	1	1	–	1	1
Brain	1	1	–	1	1
Spinal cord	1	1	–	–	–
Heart	1	1	–	1	1
Kidney	1	1	–	1	1
Liver	1	1	–	1	1
Muscle	1	1	–	1	2
Stromal cell	1	2	–	–	–
Fibroblast	1	1	–	–	–
Stem cell	1	1	–	–	–
Cloaca	1	2	–	–	–
Total	60	119(6)	23	21	34

The number in parenthesis indicates that the microarray data obtained by the Mouse430A 2 array.

probe annotation data provided by Affymetrix stated that multiple probe sets were designed for the same gene, the relationship among these probe sets was unclear in terms of the data provided. To provide information to enable the interpretation of such relationships, we have re-annotated all of the probe sets based on the results of a BLAST search against transcript sequences in RefSeq, GenBank and UniGene databases, and have classified them into six categories (Table 2). This classification was based on extents of curation for each database: RefSeq database is widely accepted as a manually curated collection of sequences representing genomes and thus contains highly reliable transcript information (Pruitt *et al.*, 2007); GenBank database is a comprehensive public repository of sequences, including data from high-throughput cDNA sequencing projects, submitted by the scientific community (Benson *et al.*, 2007); UniGene database is an informational platform for partitioning GenBank sequences, including EST, into a non-redundant set of gene-oriented clusters *in silico* (Schuler, 1997). Therefore, the probe sets in category A, which corresponded to transcripts in RefSeq, were most likely to probe mature mRNAs specifically. In contrast, we observed that ~44% and 51% of probe sets in the categories B and C on Mouse430_2 and HG_U133_Plus2, respectively, were mapped on non-exonic regions (i.e. introns or downstream of the 3′UTR) of protein-coding RefSeq genes by BLAST searches against the genome sequences. Figure 2A shows the distribution of the mean hybridization signal intensities for probe sets across 119 samples in the categories A, B and C on the Mouse430_2 array. As we expected, the mean hybridization signal intensity for category A (5.56 ± 2.87) was significantly higher than that for B (4.09 ± 2.13) and C (3.6 ± 1.67). We observed similar trends with the data for the HG_U133_Plus2 array (data not shown). These results are consistent with those previously reported (Zhang *et al.*, 2005). Based on our annotation, 10 984 and 12 008 genes were targets for multiple probe sets on Mouse430_2 and HG_U133_Plus2 arrays, respectively. We calculated Pearson's correlation coefficients for the expression patterns in different samples among the probe sets to which we assigned the same gene in a pair-wise fashion, and confirmed that there was a better correlation for pairs of probe sets from category A compared to pairs from the other categories (Fig. 2B).

We emphasize that our annotation of the probe sets provides the information, whether or not mature mRNA could be probed, for comparison of mRNA and protein profiles.

Table 2. Annotation and classification of probe sets for Affymetrix GeneChip microarrays

Category	Mouse430_2		HG-U133_Plus2	
	Number of probe sets	Number of genes	Number of probe sets	Number of genes
A	23 294	15 900	26 328	16 860
B	8881	7105	6430	5119
C	7231	4926	11 077	6115
D	1202	–	2047	–
E	1146	–	1127	–
X	3283	–	7604	–
Total	45 037	20 623	54 613	19 300

Table 3. Statistics for 2-DE gel-based proteomic analysis

Cell type	Cell subset	Number of spots quantified	Number of protein spots identified
B lymphocyte	From spleen, resting	1390	291
	From spleen, stimulated with BAFF, 7 h	1181	289
	From spleen, stimulated with CD40L and anti-CD8, 7 h	1279	288
	From spleen, stimulated with anti-IgM, 7 h	1124	287
	From spleen, stimulated with LPS, 7 h	1067	287
	CH12 cell-line, resting	1255	317
	CH12 cell-line, stimulated with anti-CD40, IL4 and TGFb1, 8 h	1256	297
	CH12 cell-line, stimulated with anti-CD40, IL4 and TGFb1, 48 h	1281	293
	WEHI-231 cell-line, resting	1257	285
	WEHI-231 cell-line, stimulated with anti-IgM, 8 h	1126	282
	WEHI-231 cell-line, stimulated with anti-IgM, 24 h	1228	282
	WEHI-231 cell-line, stimulated with anti-IgM and anti-CD40, 8 h	1223	285
	WEHI-231 cell-line, stimulated with anti-IgM and anti-CD40, 24 h	1060	283
T lymphocyte	From spleen, Th1	1003	316
	From spleen, Th2	981	284
	23-1-8 cell-line (Th1 clone)	1055	270
	MS-SB cell-line (Th2 clone)	963	269
Macrophage	J774 cell-line, resting	1347	420
	J774 cell-line, stimulated with LPS, 2 h	1494	303
Dendritic cell	From bone marrow, resting	1468	369
	From bone marrow, stimulated with LPS, 4 h	1399	292
Epithelial cell	From small intestine	990	351
	From Peyer's patches	1074	269

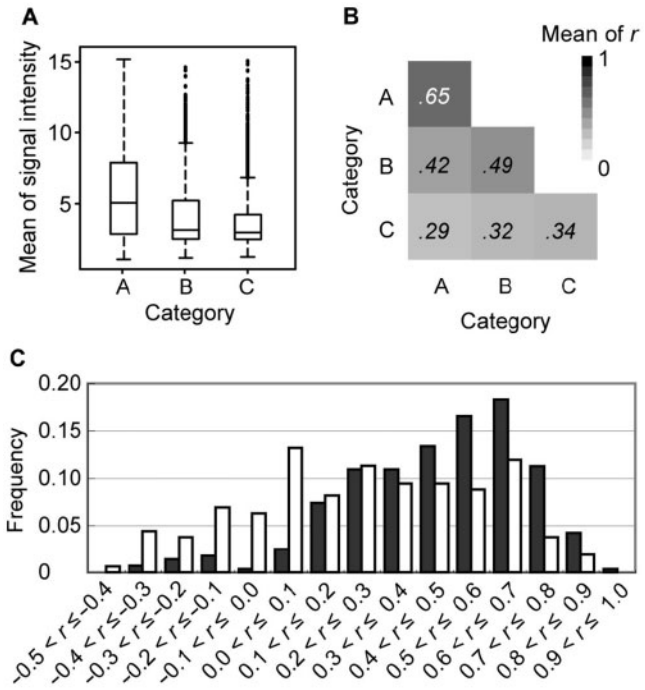

Fig. 2. (**A**) Box and whisker plot of mean hybridization signal intensities for probe set categories A, B and C on Mouse430_2 across 119 samples. The expression level for each was calculated using the gcRMA algorithm. (**B**) Similarity of expression patterns between probe-set pairs to which the same gene was assigned, though into different categories. The numerals in each element of the triangular matrix represent the average Pearson's correlation coefficients (r) for the expression patterns across the samples between the probe-set pairs. (**C**) Distribution of Pearson's correlation coefficients between the mRNA and protein levels in the 166 genes when the probe sets in category A were used (black bar) and when those in categories B or C were used (white bar).

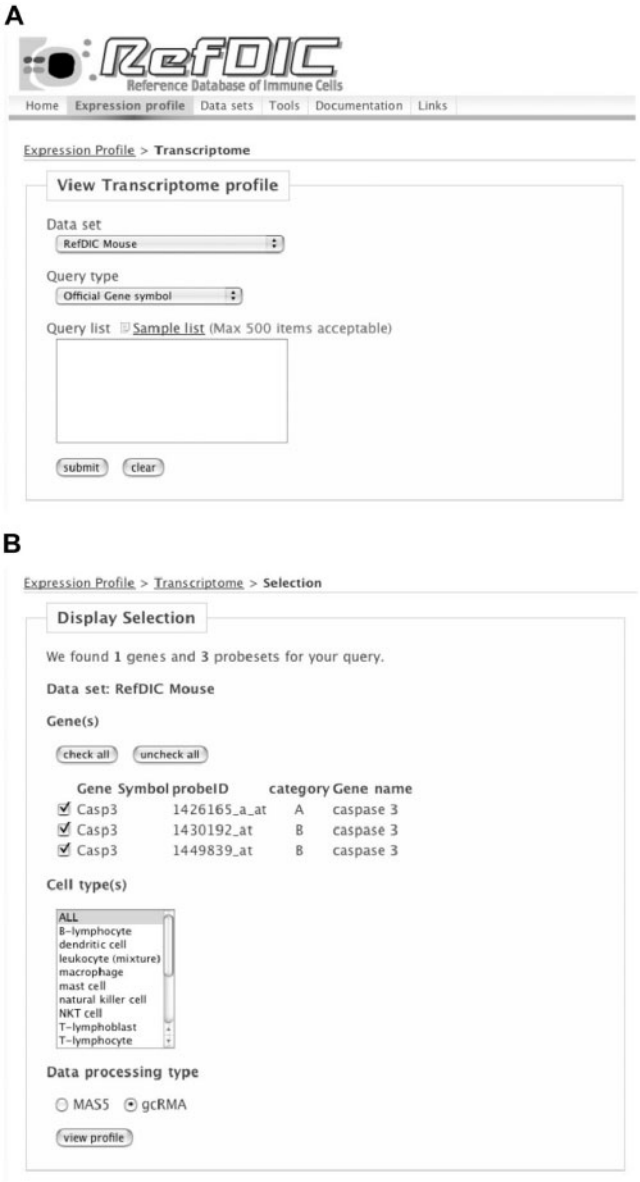

Fig. 3. (**A**) A query form for visualization of mRNA profiling data. Users can search and retrieve the mRNA profiling data for genes of interest with various types of query (i.e. gene symbol, gene/protein name, accession number of mRNA or protein sequences, GO terms or Pfam domain name). (**B**) When the users submit a query, the RefDIC server returns the list of genes available for visualization. Users can select the genes/probe sets, cell types or data processing type that they want to view.

As expected, we found that the correlation between mRNA and protein levels for 166 genes in which products could be identified in all 2-DE gel samples were improved when the probe sets in the A category were used for this calculation: the mean correlation coefficient was 0.47 ± 0.25, whereas that for categories B or C was 0.26 ± 0.31 (Fig. 2C and Supplementary Table S2). Nevertheless, we found that 61% (98 out of 159) genes have significant correlation ($P < 0.01$), when we chose category A probe sets for the calculation. The rate of occurrence of genes having significant correlation in this study was slightly higher than previously reported (Kwong *et al*., 2005; Rautajoki *et al*., 2004).

4.3 Visualization of the mRNA and protein profiling data on RefDIC

The RefDIC website and database were designed to facilitate exploration of mRNA and protein levels in immune cells on demand. This provides a query interface to enable the visualization of mRNA and protein profiling data in the following three ways:

(1) RefDIC enables the visualization of mRNA profiling data for multiple genes. One can submit a query with various types of inputs, such as gene symbol, gene/protein names and probe set ID, in order to retrieve the mRNA profiling data from the database (Fig. 3A). When a user submits a query, the RefDIC server returns a list of genes and probe sets matching the query (Fig. 3B). The list has annotation information including the probe set category, and users can select a set of genes, cell types and data processing type. The server then returns a heat map

Fig. 4. (**A**) A snapshot of the mRNA profile view. The mRNA levels for the probe sets are represented as a heat map of the different mouse samples. The top and second rows of the heat map represent the types of cell or organ, respectively. The red and green colors in the heat map show high and low mRNA levels, respectively. In this case, three probe sets (1426165_a_at, 1430192_at and 1449839_at) are mapped to caspase 3. When users click the probe set name shown this figure, they can view the genomic positions of these probe sets (inset). (**B**) Two examples of visualization of mRNA and protein levels: ornithine aminotransferase (left panel) and peroxiredoxin 3 (right panel) are shown. The levels of both mRNA and protein are represented by the relative values in log2 space.

of the mRNA profiling data that the user has selected (Fig. 4A). The profile data can be downloaded as tab-delimited text files. This viewer provides links to the chromosome map where the probe sets mapped (inset, Fig. 4A), and it provides lines of information for interpretation when the mRNA levels are discordant among the probe sets to which the same gene was assigned. The viewer also provides several links to a description of the source of samples or genes.

(2) RefDIC enables the visualization of protein profiling data for multiple genes, as in the case of the mRNA profiling data. For the protein profile viewer it also provides links to lines of information regarding the protein spot, including apparent isoelectric point,

apparent molecular mass, the mass spectral analysis, the reliability of the protein identification and the enlarged gel image (Supplementary Fig. S2A). The server also provides an interactive viewer for a whole 2-DE gel image of a particular sample accessible from '2DE-Gel image viewer' (http://refdic.rcai.riken.jp/2Dgel_viewer.cgi), and enables the user to view and search the information regarding the protein spots by clicking each protein spots on the gel image (Supplementary Fig. S2B).

(3) RefDIC enables the visualization of both mRNA and protein profiling data for a single gene, as shown in Figure 4B. In this viewer both mRNA and protein levels are represented by the relative amount of each in different samples, and thus the user can easily find at a glance whether their profiles correlate or not. As an example, Figure 4B shows that the ornithine aminotransferase gene demonstrated good correlation between mRNA and protein levels ($r^2 = 0.88$), whereas peroxiredoxin 3 showed poor correlation ($r^2 = 0.01$).

4.4 Links to the external microarray data on public repositories

To function as a data-sharing platform for genomics information in immunology, RefDIC has a function to search and retrieve the microarray experimental data in the public repositories GEO and Array Express. At present, data searching is confined to data sets using Affymetrix GeneChip Mouse430_2, Mouse430A_2 and HG_U133_Plus2 arrays simply because they are directly comparable with our data. This function would greatly facilitate data-sharing processes in immunology. In the future, we plan to add a tool to enable the visualization of expression profiles for publicly available microarray data that is relevant to immunology, together with our data on the website.

5 CONCLUSION

We have successfully developed a web-based platform for the integration of quantitative immunological transcriptomic and proteomic data. RefDIC was implemented with user-friendly query interfaces and enabled us to search and visualize both mRNA and protein profiling data for various types of immune cells at a glance. All of the information and data, including experimental raw data, are freely available via the Internet. Because this informational platform for cross-referencing data has an expandable structure, it can easily incorporate public data. RefDIC continues to grow in terms of the quantity of mRNA and protein profiling data available, and it continues to integrate more relevant information for genes and proteins, such as *cis*-elements including promoter sequences, and data for protein–protein interactions.

ACKNOWLEDGEMENTS

The authors thank Dr Tomoko Mori, Sachiko Matsuyama, Masako Mori, Shoji Tane and Yuri Ishizu for their help in performing microarray and 2-DE experiments. The authors also thank Yasuaki Murahashi and Miho Izawa for their help in setting up the web server. The first author is grateful to Advanced Center of Computing and Communication, RIKEN for kindly providing the necessary computing resources.

Conflict of Interest: none declared.

REFERENCES

Abbas,A.R. *et al.* (2005) Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes Immun.*, **6**, 319–331.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Benson,D.A. *et al.* (2007) GenBank. *Nucleic Acids Res.*, **35**, D21–D25.

Cutler,P. (2003) Protein arrays: the current state-of-the-art. *Proteomics*, **3**, 3–18.

Dai,M. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, e175.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Fievet,J. *et al.* (2004) Assessing factors for reliable quantitative proteomics based on two-dimensional gel electrophoresis. *Proteomics*, **4**, 1939–1949.

Gustafsson,J.S. *et al.* (2004) Statistical exploration of variation in quantitative two-dimensional gel electrophoresis data. *Proteomics*, **4**, 3791–3799.

Harbig,J. *et al.* (2005) A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res.*, **33**, e31.

Hubbell,E. *et al.* (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**, 1582–1592.

Irizarry,R.A. *et al.* (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.

Kersey,P.J. *et al.* (2004) The international protein index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.

Kimura,Y. *et al.* (2006) Construction of quantitative proteome reference maps of mouse spleen and lymph node based on two-dimensional gel electrophoresis. *Proteomics*, **6**, 3833–3844.

Komor,M. *et al.* (2005) Transcriptional profiling of human hematopoiesis during in vitro lineage-specific differentiation. *Stem Cells*, **23**, 1154–1169.

Kwong,K.Y. *et al.* (2005) Synchronous global assessment of gene and protein expression in colorectal cancer progression. *Genomics*, **86**, 142–158.

Maglott,D. *et al.* (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D33.

Mahon,P. *et al.* (2001) Quantitative and reproducible two-dimensional gel analysis using Phoretix 2D Full. *Electrophoresis*, **22**, 2075–2085.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D5–D12.

Rautajoki,K. *et al.* (2004) Proteome characterization of human T helper 1 and 2 cells. *Proteomics*, **4**, 84–92.

Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.

Smalheiser,N.R. (2002) Informatics and hypothesis-driven research. *EMBO rep.*, **3**, 702.

Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.

Zhang,J. *et al.* (2005) Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. *Genomics*, **85**, 297–308.

Zheng,C. *et al.* (2006) Gene expression profiling of CD34+ cells identifies a molecular signature of chronic myeloid leukemia blast crisis. *Leukemia*, **20**, 1028–1034.