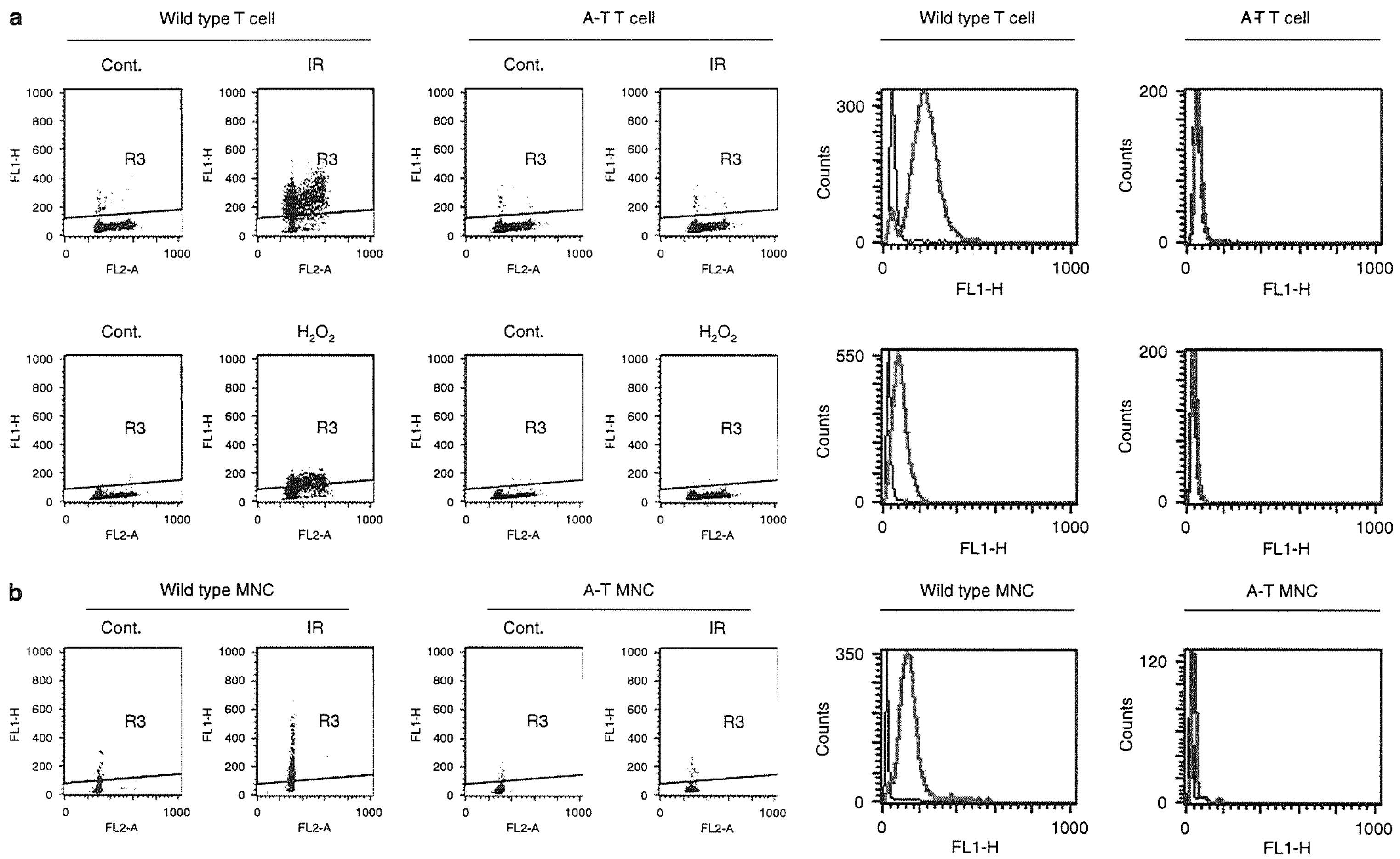

Figure 2 ATM phosphorylation status in wild-type and A-T patient. (**a**) Flowcytometric monitoring of phosphorylated ATM in activated T cells from wild-type or A-T patient. Right histogram panels correspond to the results from dot blot gram. Black line: control, non-treatment; red line: irradiated or H_2O_2 treated cells. (**b**) Flowcytometric monitoring of phosphorylated ATM from a wild-type or A-T patient without T cell stimulation. Peripheral blood mononuclear cells (PBMNCs) from a wild-type or an A-T patient were subjected to flow cytometric analysis without activation. Right histogram panels correspond to the results from dot blot gram. Black line: control, non-treatment; red line: irradiated cells.

lymphocytes to analyze ATM activation starting from a small amount of PBMNC. Then, flow cytometric analysis for ATM phosphorylation was employed in wild-type and A-T-derived activated T cells using the same protocol as that for EBV-transformed LCLs. As a result, activated T cells from wild-type control showed substantial increase of phosphorylated ATM after 10 Gy irradiation and 0.5 mM H_2O_2 treatment. On the other hand, A-T patient-derived activated T cells, which respond to IL-2 and CD3 as much as those of wild-type control (data not shown), showed no ATM phosphorylation (Figure 2a). Then, it was investigated whether this assay can be applied to PBMNC without stimulation. PBMNC from wild-type or A-T patient was isolated using Percoll gradient and irradiated soon after isolation in RPMI medium with 10% fetal calf serum. At 30 min after irradiation, PBMNC was assessed by flow cytometric analysis for ATM phosphorylation. Wild-type PBMNC showed increased phosphorylation of ATM after irradiation but no phosphorylation was detected in A-T-derived PBMNC (Figure 2b). These results suggested that flow cytometric analysis for ATM phosphorylation can be applied for the bed-side diagnosis of A-T patients even without precultivation with IL-2 and anti-CD3 antibody.

Next, this assay was tested to see whether the obligate A-T heterozygous carrier is distinguishable from a wild-type individual. Two independent obligate A-T heterozygote-derived EBV-LCLs (LCL-Hetero.), 227RM (7517 del 4/wt) and 373 RM (8283 del TC/wt), were enrolled in this study. Both of these LCLs showed increased ATM phosphorylation. However, the intensity of phosphorylation level was rather modest when compared with LCL-WT (Figure 3a). The phosphorylation level in obligate A-T heterozygotes correlated with the modest level of ATM protein detected by the western blotting method (Figure 3b).

Diagnosis of A-T is made on the basis of clinical symptoms, including ataxia and telangiectasia and the support of laboratory data. Although the demonstration of lack of ATM protein by western blotting method is a straightforward approach for the screening of A-T patients, ATM protein is not expressed enough to identify wild-type controls in PBMNC, thus making it difficult to distinguish normal, heterozygous and homozygous individuals. Establishment of cell lines or stimulation of PBMNC is required for western blotting analysis. Protein truncation assay is not practical for clinical use because of the complexity of the technique and requirement of several special equipment. Furthermore, these methods failed to identify the missense mutations with normal expression levels of mutant ATM. Sequence-based genetic diagnosis is not easy because of the large size of the *ATM* gene. The rapid screening methods that precisely predict the presence of ATM mutation will save effort, time and cost for the selection of patients who need whole ATM gene sequencing. Under the light of these previous laboratory experiences, we have eagerly awaited the development of a convenient screening procedure. Here we developed a method to detect the expression levels of phosphorylated ATM protein, which is based on the combinational analyses of ATM protein and its phosphorylation levels. This functional analysis is expected to enable screening of genomic condition of A-T gene before

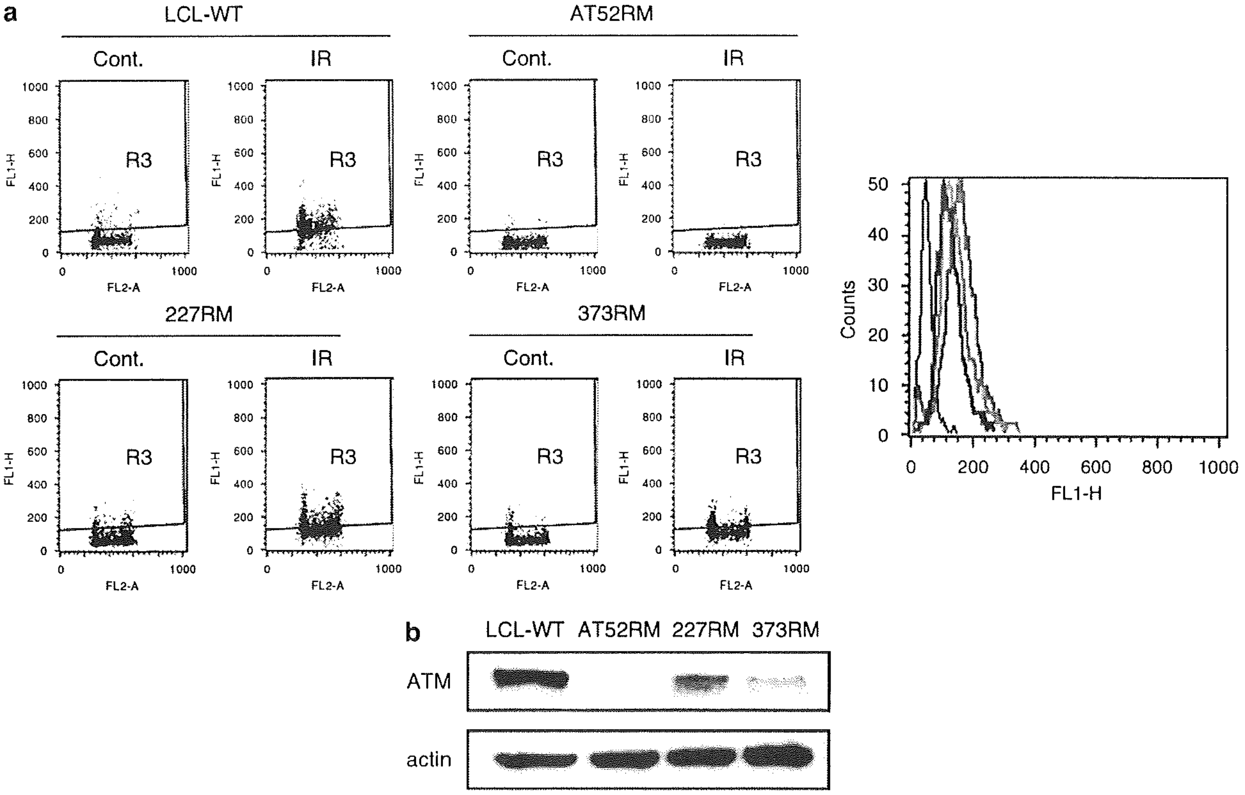

Figure 3 ATM phosphorylation status in A-T heterozygous carrier derived LCLs in comparison with LCL-WT and AT52RM with ATM deficiency. (**a**) ATM phosphorylation was measured 1 h after 10 Gy irradiation. Right histogram panel corresponds to the results from dot blot gram. Black line: AT52RM; blue line: 227RM; green line: 373RM; red line: LCL-WT. The histogram illustrates ATM phosphorylation after irradiation of each cell line. The data from irradiated samples is shown. The histogram from un-irradiated sample of each LCL overlapping with that of irradiated AT52RM was omitted. (**b**) western blotting data of ATM protein expression from LCL-WT, AT52RM and LCL-Hetero. (227RM, 373RM).

genome sequencing. Flow cytometry is well equipped in most of the clinical laboratories, is relatively easy to manipulate and is easily accessible. Once the procedure is developed, it enables us to share this method between basic and clinical researchers.

The results on EBV-LCLs further suggest that this method will facilitate the diagnosis of heterozygous carriers. Thus, an advantage of our flow cytometric analysis using ATM phosphorylation also exists in the diagnosis of carriers with missense mutations. It was shown that overexpression of dominant-negative form of misssense mutation at the kinase domain interferes with the ATM autophosphorylation after DNA damage. Further investigation is underway to find whether we can identify the individuals with obligatory heterozygous missense mutation by this assay without EBV transformation.

Flow cytometric analysis enables us to screen simultaneously relatively large numbers of patient samples. We have previously reported that activation of DNA damage checkpoints plays an important role in the prevention of leukemic transformation in myelodysplastic syndrome. Phosphorylation of ATM is detected in refractory anemia with excess blast (RAEB)-II phase but not in RA phase.[7] It is often difficult to morphologically classify these two conditions by microscopy. Measurement of ATM phosphorylation using flow cytometry might help to subclassify myelodysplastic syndrome. Furthermore, acquired mutation of *ATM* gene is frequently identified in hematological malignancies, such as T-prolymphocytic leukemia, mantle cell lymphoma, B-chronic lymphocytic leukemia. Loss of ATM function shows poorer clinical outcome in B-chronic lymphocytic leukemia.[8] The classification of risk groups using ATM activation may add one of the options for selecting the therapy. As with the other clinical applications, measurement of ATM phosphorylation using flow cytometry may be useful for identification of individuals who may show abnormal response against genotoxic stress such as exposure to radiation or chemotherapeutic agents. If these individuals are screened at the bedside, we can avoid the adverse effect of anticancer drugs by reducing the administration doses. Additionally, this screening method will extend the opportunities for risk assessment of cancer development on the basis of population by identifying heterozygous carriers of the *ATM* gene in individuals with or without cancer.

Acknowledgements

This work was supported by a Grant-in-Aid from the Ministry of Education, Science, and Culture, Japan, by a Grant-in-Aid for Cancer Research from the Ministry of Health, Labor and Welfare, Japan, and by a Grant-in-Aid for Child Health and Development from the Ministry of Health, Labor and Welfare, Japan.

M Honda[1], M Takagi[1], L Chessa[2], T Morio[1] and S Mizuatni[1]
[1]*Department of Pediatrics and Developmental Biology, Graduate School of Medicine, Tokyo Medical and Dental University, Bunkyo-ku, Tokyo, Japan and*
[2]*Department of Experimental Medicine Pathology, II Faculty of Medicine, University 'La Sapienza', Rome, Italy*
E-mails: m.takagi.ped@tmd.ac.jp or smizutani.ped@tmd.ac.jp

References

1 Savitsky K, Sfez S, Tagle DA, Ziv Y, Sartiel A, Collins FS *et al.* The complete sequence of the coding region of the ATM gene reveals similarity to cell cycle regulators in different species. *Hum Mol Genet* 1995; **4**: 2025–2032.

2 Lavin MF, Gueven N, Bottle S, Gatti RA. Current and potential therapeutic strategies for the treatment of ataxia-telangiectasia. *Br Med Bull* 2007; **81–82**: 129–147.

3 Stankovic T, Stewart GS, Byrd P, Fegan C, Moss PA, Taylor AM. ATM mutations in sporadic lymphoid tumours. *Leuk Lymphoma* 2002; **43**: 1563–1571.

4 Shiloh Y. ATM and related protein kinases: safeguarding genome integrity. *Nat Rev Cancer* 2003; **3**: 155–168.

5 Bakkenist CJ, Kastan MB. DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. *Nature* 2003; **421**: 499–506.

6 Sekine T, Shiraiwa H, Yamazaki T, Tobisu K, Kakizoe T. A feasible method for expansion of peripheral blood lymphocytes by culture with immobilized anti-CD3 monoclonal antibody and interleukin-2 for use in adoptive immunotherapy of cancer patients. *Biomed Pharmacother* 1993; **47**: 73–78.

7 Horibe S, Takagi M, Unno J, Nagasawa M, Morio T, Arai A *et al.* DNA damage check points prevent leukemic transformation in myelodysplastic syndrome. *Leukemia* 2007; **21**: 2195–2198.

8 Austen B, Skowronska A, Baker C, Powell JE, Gardiner A, Oscier D *et al.* Mutation status of the residual ATM allele is an important determinant of the cellular response to chemotherapy and survival in patients with chronic lymphocytic leukemia containing an 11q deletion. *J Clin Oncol* 2007; **25**: 5448–5457.

Expression of cyclin A in childhood acute lymphoblastic leukemia cells reveals undisturbed G1–S phase transition and passage through the S phase

Leukemia (2009) **23**, 414–417; doi:10.1038/leu.2008.200; published online 31 July 2008

Pediatric acute lymphoblastic leukemia (ALL) is characterized by arrested differentiation of the malignant cell clone with retention of proliferative ability. Earlier we had demonstrated that the majority of leukemic cells reside in the G1 phase of the cell cycle, with a minor proportion of cells in the S and G2 phases and only a few cells in the true G0 phase.[1] Although the majority of ALL cells reside in the G1 phase of the cell cycle, attempts to induce differentiation *in vivo* or *in vitro* have not been successful.

In view of this arrested differentiation, we became interested in the distribution of ALL cells within the G1 phase relative to the restriction (R) point. The R point determines whether a cell is committed to proliferation or retains the ability to differentiate.[2] Progression through the R point is regulated by the functional inactivation of the retinoblastoma protein (pRb) through a complex stepwise process that includes a sequential phosphorylation by at least two different cyclin-dependent kinase (cdk) complexes, cyclin D/cdk4,6 and cyclin E/cdk2 inducing a loss of nuclear tethering and the activation of E2F transcriptional programs that permit G1–S phase transition.[3]

In pediatric ALL cell samples, both cyclin D/cdk4,6 and cyclin E/cdk2 kinase complexes were active; correspondingly,

Prediction of Candidate Primary Immunodeficiency Disease Genes Using a Support Vector Machine Learning Approach

Shivakumar Keerthikumar[1,2,3], Sahely Bhadra[4], Kumaran Kandasamy[1,2,5], Rajesh Raju[1,2,3], Y.L. Ramachandra[2], Chiranjib Bhattacharyya[4], Kohsuke Imai[8], Osamu Ohara[6,7], Sujatha Mohan[1,3], and Akhilesh Pandey[1,5,*]

Institute of Bioinformatics, International Technology Park, Bangalore 560 066, India[1]; Department of Biotechnology and Bioinformatics, Kuvempu University, Jnanasahyadri, Shimoga 577 451, India[2]; Research Unit for Immunoinformatics, Research Center for Allergy and Immunology, RIKEN Yokohama Institute, Kanagawa 230-0045, Japan[3]; Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560 012, India[4]; McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, 733 N. Broadway, BRB Room 527, Baltimore, MD 21205, USA[5]; Laboratory for Immunogenomics, Research Center for Allergy and Immunology, RIKEN, Yokohama Institute, Kanagawa 230-0045, Japan[6]; Department of Human Genome Technology, Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan[7] and Department of Medical Informatics, National Defense Medical College, Saitama 359-8513, Japan[8]

(Received 23 July 2009; accepted 5 September 2009; published online 3 October 2009)

Abstract

Screening and early identification of primary immunodeficiency disease (PID) genes is a major challenge for physicians. Many resources have catalogued molecular alterations in known PID genes along with their associated clinical and immunological phenotypes. However, these resources do not assist in identifying candidate PID genes. We have recently developed a platform designated Resource of Asian PDIs, which hosts information pertaining to molecular alterations, protein–protein interaction networks, mouse studies and microarray gene expression profiling of all known PID genes. Using this resource as a discovery tool, we describe the development of an algorithm for prediction of candidate PID genes. Using a support vector machine learning approach, we have predicted 1442 candidate PID genes using 69 binary features of 148 known PID genes and 3162 non-PID genes as a training data set. The power of this approach is illustrated by the fact that six of the predicted genes have recently been experimentally confirmed to be PID genes. The remaining genes in this predicted data set represent attractive candidates for testing in patients where the etiology cannot be ascribed to any of the known PID genes.

Key words: RAPID; SVM; HPRD; Human Proteinpedia; NetPath

1. Introduction

Primary immunodeficiency diseases (PIDs) are a genetically heterogeneous group of disorders that affect distinct components of the innate and adaptive immune system, such as neutrophils, macrophages, dendritic cells, natural killer cells and T and B lymphocytes. The study of these diseases has provided essential insights into the functioning of our immune system. More than 120 distinct genes have been identified, whose abnormalities account for more than 150 distinct forms of PID.[1] PIDs are challenging for both researchers and clinicians because they represent natural models of immunopathology, which can usually be studied effectively only in animal models, and manifest with a wide range of clinical symptoms ranging from susceptibility to infections

Edited by Minoru Ko

* To whom correspondence should be addressed. Tel. +1 410-502-6662. Fax. +1 410-502-7544. E-mail: pandey@jhmi.e.du

and allergies to autoimmune and inflammatory diseases. The genetic defects that cause PIDs can affect the expression and function of proteins involved in a range of biological processes, such as immune development, effector-cell functions, signaling cascades and maintenance of immune homeostasis.[2]

Because genes and proteins rarely work in isolation, genes that directly or functionally interact with known PID genes could also represent additional PID genes. We have recently developed a database of PID genes designated 'Resource of Asian PDIs (RAPID)', which contains information pertaining to genes and proteins involved in PDIs along with other relevant information about protein–protein interactions, mouse knockout studies and microarray gene expression profiles in various cells and organs of the immune system. These significant features of PID genes, including their involvement in immune signaling pathways, were used as input binary features for the prediction of additional candidate PID genes using a support vector machine (SVM) learning approach.

SVM is a powerful machine learning technique widely used in the computational biology such as microarray data analysis,[3–8] protein secondary structure prediction,[9] prediction of human signal peptide cleavage sites,[10] translational initiation site recognition in DNA,[11] protein fold recognition,[12,13] prediction of protein–protein interactions,[14] prediction of protein sub-cellular localization,[15–18] and peptide identification from mass-spectrometry derived data.[19]

SVM is a learning algorithm that can be used to generate a classifier from a set of positively and negatively labeled training data sets.[20] SVM learns the classifier by mapping the input training samples into a possibly high-dimensional feature space and seeking a hyperplane in this space, which separates the two types of examples with the largest possible margin, i.e. distance to the nearest points. If the training set is not linearly separable, SVM finds a hyperplane, which optimizes a trade-off between good classification and large margin.[20]

For predicting a classifier between PID and non-PID genes, we have solved the above problem and obtained a linear classifier (Fig. 1). To prove generalization of the predicted classifier, we have reported leave-one-out (LOO) error for the training data set. In this approach, we have used all the known PID genes that have been described in the literature as a positive data set. The gene list for negative data sets was selected from mouse genomic informatics (MGI) database based on the criterion that mutations in mice do not result in either immune or hematopoietic system phenotypes. We trained SVM with 69 features (Supplementary Table S1) for both PID genes (positive data set) and genes that were not reported to be associated with PIDs (negative data set). The trained SVM was then used to predict candidate PID genes by testing all human genes (except those used in the training data sets) as test data set.

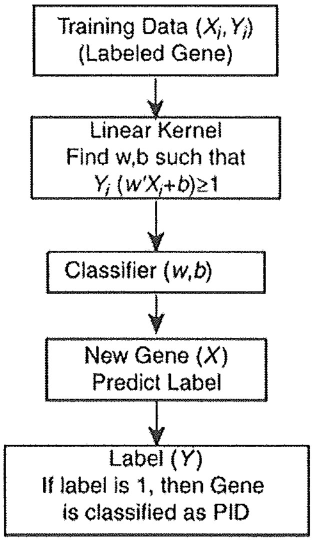

Figure 1. A schematic of SVM training strategy.

2. Materials and methods

2.1. *Initial platform*

RAPIDs, which is available as a worldwide web resource at http://rapid.rcai.riken.jp/[21] was used as a source of information about PID genes. RAPID hosts information on sequences and expression at the mRNA and protein levels of genes reported to be involved in PID patients. The main objective of this database was to provide detailed information pertaining to genes and proteins involved in PIDs along with other relevant information about protein–protein interactions, mouse knockout studies and microarray gene expression profiles in various organs and cells of the immune system.

2.2. *Features used for training the data sets*

The PDIs are characterized by essential defects in the functions of the immune system, leading to increased susceptibility to infections. Although rare, these disorders cover a wide spectrum of defects, including antibody deficiencies, cellular immune deficiencies, combined immune deficiencies, phagocytic defects, complement and other innate immunity defects. On the basis of these observations for all the known PID genes, we selected 69 features (Supplementary Table S1) which not only play an important role in the development, maintenance and normal functioning of immune/hematopoietic systems but also in understanding molecular pathophysiology of PID disease causing genes. These features can be broadly classified as features for signaling pathways from NetPath and KEGG[22–24] database, microarray gene expression profile from RefDIC[25] database, site of expression from HPRD[26] and Human Proteinpedia,[27] immune/hematopoietic phenotypes from MGI[28,29] and interaction with PID feature from HPRD.

2.3. *Data sets*

To train the SVM, two types of data sets were generated—the positive data set consists of all the known PID genes, whereas the negative data set contained genes where no immune/hematopoietic system abnormalities were described due to mouse knockouts, knockins or spontaneous mutations reported for the mouse orthologs in the MGI database.[30] On the basis of these criteria, 148 PID genes were in the positive data set and 3162 genes were in the negative data set. Test data set contains 36 677 genes encoded by the human genome. Genes involved in both the training and test data sets were assigned a binary score of '1' and '0,' respectively, based on their presence or absence in a particular feature. The trained SVM was used to classify PID or non-PID genes from an unlabeled test data set which consists of all human genes (Fig. 2).

2.4. *SVM implementation*

We used SVM^{light} (http://svmlight.joachims.org/), an implementation of Support Vector Machines in C, and also used customized functions written in MATLAB (http://www.mathtools.net/MATLAB/) for the calculation of confidence score for each predicted candidate PID gene. Absolute score also known as confidence score can be defined as AbsScore(X) = ($w^{\mathrm{T}}X - b$) where $w^{\mathrm{T}}x - b = 0$ represents the separating hyperplane calculated by SVM. The score indicates how far that particular gene from the positive side of the hyperplane. In other words, higher the score more likely that a particular gene is a candidate PID gene. Using this approach, 1442 candidate PID genes were predicted which falls on the positive side of the hyperplane.

2.5. *LOO error*

LOO error measurement involves removing one gene from the training set, training the SVM on the remaining genes and then predicting the class label of that gene that was left out. This process is repeated until all the genes are left out exactly once. If the gene was classified correctly, the error was reported as zero,

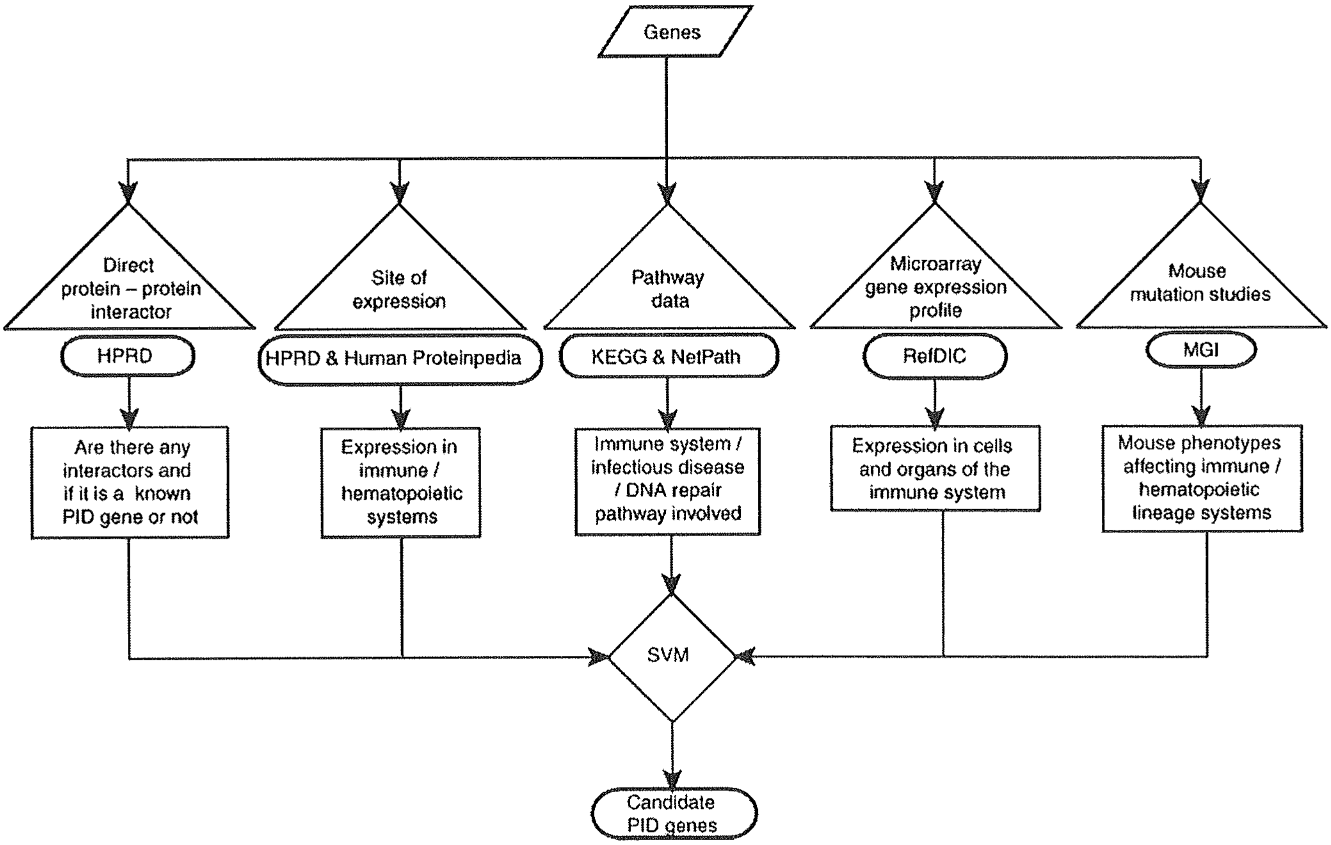
Genes
Direct protein – protein interactor
Site of expression
Pathway data
Microarray gene expression profile
Mouse mutation studies
HPRD
HPRD & Human Proteinpedia
KEGG & NetPath
RefDIC
MGI
Are there any interactors and if it is a known PID gene or not
Expression in immune / hematopoietic systems
Immune system / infectious disease / DNA repair pathway involved
Expression in cells and organs of the immune system
Mouse phenotypes affecting immune / hematopoietic lineage systems
SVM
Candidate PID genes

Figure 2. A schematic of the algorithm for prediction of candidate PID genes.

else the error was reported as one. This process was repeated by leaving out each gene once and the LOO error of the data set represent the average of individual errors.

3. Results and discussion

Over 1500 Mendelian disorders whose molecular basis is unknown are catalogued in the online Mendelian inheritance in man (OMIM) database.[31] Most of disease-gene identification efforts involve either linkage analysis or association studies.[32,33] Recently, a number of *in silico* approaches to identify candidate disease genes have been developed that use available information reported from various studies such as functional annotation, gene expression profiles, annotated sequence features, protein–protein interactions and pathway information.[34–39] Several machine learning approaches have also been employed to identify important genes for disease classification. SVM approach is generally preferred owing to its superior performance.[40] In most instances, SVM is a powerful tool in dealing with high-dimensional low sample size data sets, which also performs well in various biological analyses including text categorization, evaluating microarray expression data and inferring functional annotation from protein sequence and structure data.[3,4,41,42] In this study, we trained an SVM with 69 features for both positive (all known PID genes) and negative (genes with no immune/hematopoietic systems affected due to mutations from MGI) gene data sets.

As the number of genes in the positive data set is small, the LOO error was calculated for showing generalization of the algorithm. LOO error is explained in detail under the Materials and methods section. For this, we used a data set containing 148 PID genes from positive data sets along with 148 genes that were randomly selected from the negative data set. This process was repeated and from 60 such data sets, the LOO error was calculated. The average LOO error reported over 60 data sets was ~8%. The LOO error reported by leaving out only the PID (positive) genes one by one (where training set contains same setting of 296 data points) was ~15%.

3.1. *Sensitivity and specificity*

The sensitivity and specificity of the data sets was 0.85 and 0.98, respectively. On the basis of these results, we conclude that the number of genes falsely predicted to be PID genes by the trained classifier is ~2%. We believe that availability of comprehensive and accurate biological data is a limitation that restricts the prediction accuracy and performance of this algorithm. As more data accumulates about the human genome and proteome, we expect the performance of this algorithm to improve further in the future. The complete list of predicted candidate genes is provided in Supplementary Table S2 and also available at the RAPID website http://rapid.rcai.riken.jp/. All 69 features of the predicted candidate PID genes can also be downloaded from the RAPID website.

3.2. *Evaluation studies*

We were able to evaluate our predictions in a limited fashion because a few studies have been published describing novel PID genes that were not included in our original list of PID genes. These experimental studies have confirmed six of the genes in our predicted list of PID genes as true PID genes. These are myeloid differentiation factor-88 (*MYD88*), catalytic subunit of DNA dependent serine/threonine protein kinase (*PRKDC*), glucose-6-phosphatase, catalytic subunit 3 (*G6PC3*),[43–45] IL2-inducible T-cell kinase (*ITK*), coronin, actin binding protein 1A (*CORO1A*) and Interleukin 1 receptor antagonist (*IL1RN*).[46–49] MyD88 is a key downstream adaptor protein in IL1 receptor complex and toll-like receptors signaling pathways involved in inflammatory response and host defense. In addition, MyD88 is also involved in tumorigenesis in models of hepatocarcinoma and familial associated polyposis; negative regulation of TLR3 signaling and in PKC epsilon activation.[50] Patients with MyD88 deficiency are reported to be susceptible to pyogenic bacterial infections including invasive pneumococcal disease.[45] Defect in *PRKDC* has been reported for the first time in a radiosensitive T-B-SCID patient that results in inhibition of Artemis activation and non-homologous end-joining.[44] A report of mutations in *G6PC3* gene has been observed among patients with severe congenital neutropenia syndrome and also shown to be susceptible to increased apoptosis that leads to disturbances in cardiac or urogenital development.[43] A novel PDI, IL-2 inducible T-cell kinase (ITK) deficiency has been observed due to fatal immune dysregulation followed by EBV infection and identified homozygous mutation in the SH2 domain of ITK gene that resulted in protein destabilization and absence of NKT cells.[47] A patient with T cell-deficient, B cell-sufficient and NK cell-sufficient severe combined immunodeficiency has been identified with mutation in CORO1A gene along with reduced T-cell function that was earlier demonstrated in knock-out mice of coro1a gene with similar phenotypes.[49] Deficiency of the IL1-receptor antagonist, an autosomal recessive autoinflammatory disease, has been reported for the first time in children presented with clinical phenotypes of multifocal osteomyelitis, periostitis, pustulosis, thrombosis and

Table 1. A list of predicted PID genes whose association with immunological disorders has been reported recently

Gene symbol	Molecule class	Immunological disorder(s)	Reference(s)
ITGAM	Cell surface receptor	Systemic lupus erythematosus	Harley et al., *Nat. Genet.*, 2008 (PubMed ID: 18204446);[55] Nath et al., *Nature Genetics*, 2008 (PubMed ID: 18204448)[56]
BANK1	Chaperone	Systemic lupus erythematosus	Kozyrev et al., *Nat. Genet.*, 2008 (PubMed ID: 18204447)[57]
MST1	Growth factor	Inflammatory bowel disease	Goyette et al., *Mucosal Immunol.*, 2008 (PubMed ID: 19079170)[58]
CYLD	Ubiquitin–proteasome system protein	Crohn's disease	Johnson and O'Donnell et al., *BMC Med. Genet.*, 2009 (PubMed ID: 19161620)[59]
PTPN2	Tyrosine phosphatase	Crohn's disease	Wellcome Trust Case Control Consortium, 2007;[60] Todd et al., *Nat. Genet.*, 2007 (PubMed ID: 17554260)[61]
PTPN22	Tyrosine phosphatase	Systemic lupus erythematosus	Wellcome Trust Case Control Consortium, 2007;[60] Harley et al., *Nat. Genet.*, 2008 (PubMed ID: 18204446)[55]
TNFAIP3	Transcription regulatory protein	Rheumatoid arthritis	Plenge et al., *Nat. Genet.*, 2007 (PubMed ID: 17982456)[62]
STAT4	Transcription factor	Systemic lupus erythematosus	Remmers et al., *N Engl J Med.*, 2007 (PubMed ID: 17804842)[63]
TNFSF4	Ligand	Systemic lupus erythematosus	Graham et al., *Nat. Genet.*, 2008 (PubMed ID: 18059267)[64]
CTLA4	Adhesion molecule	Autoimmune thyroid diseases	Ueda et al., *Nature*, 2003 (PubMed ID: 12724780);[65] Ikegami et al., *J. Clin. Endocrinol. Metab.*, 2006 (PubMed ID: 16352685)[66]

respiratory insufficiency due to the homozygous deletion of the IL1RN gene.[46,48] Further, functional analysis of these mutants confirmed diminished or lack of mRNA and protein expressions leading to cytokine abnormalities.

There are two recent independent reports[51,52] on the identification and prioritization of candidate disease genes in general as well as specific to primary immunodeficiencies by integrating functional annotations from gene ontology and compilation of protein interaction network data sets from BIND,[53] BioGRID[54] and HPRD.[26] In the latter studies, 24 candidate genes were reported that are likely to be involved in PID have been identified using these parameters, out of which, over 80% of these genes are already listed as candidates in our SVM analysis, thereby, paving the way for successful implementation of this approach in the future.

We have also summarized reports of genome-wide association studies and other related studies for newly identified candidate PID genes and the associated immunological disorder (Table 1). Because the candidate PID gene list is still large, this approach of integrating data from high-throughput studies would allow further prioritization of genes for confirmation in patients with PID where the exact gene is not yet identified. We hope that such integrated approaches should assist PID physicians and researchers to gain insights into the pathophysiology of these diseases at a faster pace, which could be translated to improve the diagnosis and/or treatment of PIDs.

3.3. *Availability*

The list of predicted PID genes is available as Supplementary Table S2 and at the RAPID website http://rapid.rcai.riken.jp/.

Acknowledgements: The authors thank Shigeaki Nonoyama, Hirokazu Kanegane, Toshio Miyawaki, Koichi Oshima and Atsushi Hijikata for their valuable input and suggestions.

Supplementary Data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

We thank the Department of Biotechnology of the Government of India for research support to the Institute of Bioinformatics, Bangalore.

References

1. Geha, R.S., Notarangelo, L.D., Casanova, J.L., et al. 2007, Primary immunodeficiency diseases: an update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee, *J. Allergy Clin. Immunol.*, **120**, 776–94.
2. Marodi, L. and Notarangelo, L.D. 2007, Immunological and genetic bases of new primary immunodeficiencies, *Nat. Rev. Immunol.*, **7**, 851–61.
3. Brown, M.P., Grundy, W.N., Lin, D., et al. 2000, Knowledge-based analysis of microarray gene

expression data by using support vector machines, *Proc. Natl Acad. Sci. USA*, **97**, 262–7.

4. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M. and Haussler, D. 2000, Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, **16**, 906–14.
5. Pirooznia, M., Yang, J.Y., Yang, M.Q. and Deng, Y. 2008, A comparative study of different machine learning methods on microarray gene expression data, *BMC Genomics*, **9**, Suppl 1, S13.
6. Wang, L., Zhu, J. and Zou, H. 2008, Hybrid huberized support vector machines for microarray classification and gene selection, *Bioinformatics*, **24**, 412–9.
7. Wang, Y., Tetko, I.V., Hall, M.A., et al. 2005, Gene selection from microarray data for cancer classification—a machine learning approach, *Comput. Biol. Chem.*, **29**, 37–46.
8. Yeang, C.H., Ramaswamy, S., Tamayo, P., et al. 2001, Molecular classification of multiple tumor types, *Bioinformatics*, **17**, Suppl 1, S316–22.
9. Hua, S. and Sun, Z. 2001, A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach, *J. Mol. Biol.*, **308**, 397–407.
10. Jagla, B. and Schuchhardt, J. 2000, Adaptive encoding neural networks for the recognition of human signal peptide cleavage sites, *Bioinformatics*, **16**, 245–50.
11. Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T. and Muller, K.R. 2000, Engineering support vector machine kernels that recognize translation initiation sites, *Bioinformatics*, **16**, 799–807.
12. Cai, Y.D., Liu, X.J., Xu, X. and Zhou, G.P. 2001, Support vector machines for predicting protein structural class, *BMC Bioinformatics*, **2**, 3.
13. Ding, C.H. and Dubchak, I. 2001, Multi-class protein fold recognition using support vector machines and neural networks, *Bioinformatics*, **17**, 349–58.
14. Bock, J.R. and Gough, D.A. 2001, Predicting protein–protein interactions from primary structure, *Bioinformatics*, **17**, 455–60.
15. Bhasin, M. and Raghava, G.P. 2004, ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST, *Nucleic Acids Res.*, **32**, W414–9.
16. Garg, A., Bhasin, M. and Raghava, G.P. 2005, Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search, *J. Biol. Chem.*, **280**, 14427–32.
17. Hua, S. and Sun, Z. 2001, Support vector machine approach for protein subcellular localization prediction, *Bioinformatics*, **17**, 721–8.
18. Shi, J.Y., Zhang, S.W., Pan, Q., Cheng, Y.M. and Xie, J. 2007, Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition, *Amino Acids*, **33**, 69–74.
19. Anderson, D.C., Li, W., Payan, D.G. and Noble, W.S. 2003, A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores, *J. Proteome. Res.*, **2**, 137–46.
20. Park, K.J., Gromiha, M.M., Horton, P. and Suwa, M. 2005, Discrimination of outer membrane proteins using support vector machines, *Bioinformatics*, **21**, 4223–9.
21. Keerthikumar, S., Raju, R., Kandasamy, K., et al. 2009, RAPID: resource of Asian primary immunodeficiency diseases, *Nucleic Acids Res.*, **37**, D863–7.
22. Kanehisa, M., Araki, M., Goto, S., et al. 2008, KEGG for linking genomes to life and the environment, *Nucleic Acids Res.*, **36**, D480–4.
23. Kanehisa, M. and Goto, S. 2000, KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, **28**, 27–30.
24. Kanehisa, M., Goto, S., Hattori, M., et al. 2006, From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Res.*, **34**, D354–7.
25. Hijikata, A., Kitamura, H., Kimura, Y., et al. 2007, Construction of an open-access database that integrates cross-reference information from the transcriptome and proteome of immune cells, *Bioinformatics*, **23**, 2934–41.
26. Keshava Prasad, T.S., Goel, R., Kandasamy, K., et al. 2009, Human protein reference database—2009 update, *Nucleic Acids Res.*, **37**, D767–72.
27. Kandasamy, K., Keerthikumar, S., Goel, R., et al. 2009, Human Proteinpedia: a unified discovery resource for proteomics research, *Nucleic Acids Res.*, **37**, D773–81.
28. Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A. and Richardson, J.E. 2009, The mouse genome database genotypes:phenotypes, *Nucleic Acids Res.*, **37**, D712–9.
29. Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E. and Blake, J.A. 2008, The mouse genome database (MGD): mouse biology and model systems, *Nucleic Acids Res.*, **36**, D724–8.
30. Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A. and Richardson, J.E. 2007, The mouse genome database (MGD): new features facilitating a model system, *Nucleic Acids Res.*, **35**, D630–7.
31. Amberger, J., Bocchini, C.A., Scott, A.F. and Hamosh, A. 2009, McKusick's online Mendelian inheritance in man (OMIM), *Nucleic Acids Res.*, **37**, D793–6.
32. Botstein, D. and Risch, N. 2003, Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease, *Nat. Genet.*, **33**, Suppl, 228–37.
33. Glazier, A.M., Nadeau, J.H. and Aitman, T.J. 2002, Finding genes that underlie complex traits, *Science*, **298**, 2345–9.
34. Freudenberg, J. and Propping, P. 2002, A similarity-based method for genome-wide prediction of disease-relevant human genes, *Bioinformatics*, **18**, Suppl 2, S110–5.
35. Huang, D. and Chow, T.W. 2007, Identifying the biologically relevant gene categories based on gene expression and biological data: an example on prostate cancer, *Bioinformatics*, **23**, 1503–10.
36. Kohler, S., Bauer, S., Horn, D. and Robinson, P.N. 2008, Walking the interactome for prioritization of candidate disease genes, *Am. J. Hum. Genet.*, **82**, 949–58.

37. Perez-Iratxeta, C., Bork, P. and Andrade, M.A. 2002, Association of genes to genetically inherited diseases using data mining, *Nat. Genet.*, **31**, 316–9.
38. Segal, E., Wang, H. and Koller, D. 2003, Discovering molecular pathways from protein interaction and gene expression data, *Bioinformatics*, **19**, Suppl 1, i264–71.
39. Wang, K., Li, M. and Bucan, M. 2007, Pathway-based approaches for analysis of genomewide association studies, *Am. J. Hum. Genet.*, **81**, 1278–1283.
40. Zhang, H.H., Ahn, J., Lin, X. and Park, C. 2006, Gene selection using support vector machines with non-convex penalty, *Bioinformatics*, **22**, 88–95.
41. Lewis, D.P., Jebara, T. and Noble, W.S. 2006, Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure, *Bioinformatics*, **22**, 2753–60.
42. Radivojac, P., Peng, K., Clark, W.T., et al. 2008, An integrated approach to inferring gene-disease associations in humans, *Proteins*, **72**, 1030–7.
43. Boztug, K., Appaswamy, G., Ashikov, A., et al. 2009, A syndrome with congenital neutropenia and mutations in G6PC3, *N. Engl. J. Med.*, **360**, 32–43.
44. van der Burg, M., Ijspeert, H., Verkaik, N.S., et al. 2009, A DNA-PKcs mutation in a radiosensitive T-B- SCID patient inhibits Artemis activation and nonhomologous end-joining, *J. Clin. Invest.*, **119**, 91–8.
45. von Bernuth, H., Picard, C., Jin, Z., et al. 2008, Pyogenic bacterial infections in humans with MyD88 deficiency, *Science*, **321**, 691–6.
46. Aksentijevich, I., Masters, S.L., Ferguson, P.J., et al. 2009, An autoinflammatory disease with deficiency of the interleukin-1-receptor antagonist, *N. Engl. J. Med.*, **360**, 2426–37.
47. Huck, K., Feyen, O., Niehues, T., et al. 2009, Girls homozygous for an IL-2-inducible T cell kinase mutation that leads to protein deficiency develop fatal EBV-associated lymphoproliferation, *J. Clin. Invest.*, **119**, 1350–8.
48. Reddy, S., Jia, S., Geoffrey, R., et al. 2009, An autoinflammatory disease due to homozygous deletion of the IL1RN locus, *N. Engl. J. Med.*, **360**, 2438–44.
49. Shiow, L.R., Roadcap, D.W., Paris, K., et al. 2008, The actin regulator coronin 1A is mutant in a thymic egress-deficient mouse strain and in a patient with severe combined immunodeficiency, *Nat. Immunol.*, **9**, 1307–15.
50. Kenny, E.F. and O'Neill, L.A. 2008, Signalling adaptors used by toll-like receptors: an update, *Cytokine*, **43**, 342–9.
51. Chen, J., Aronow, B.J. and Jegga, A.G. 2009, Disease candidate gene identification and prioritization using protein interaction networks, *BMC Bioinformatics*, **10**, 73.
52. Ortutay, C. and Vihinen, M. 2009, Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies, *Nucleic Acids Res.*, **37**, 622–8.
53. Bader, G.D., Betel, D. and Hogue, C.W. 2003, BIND: the biomolecular interaction network database, *Nucleic Acids Res.*, **31**, 248–50.
54. Breitkreutz, B.J., Stark, C., Reguly, T., et al. 2008, The BioGRID interaction database: 2008 update, *Nucleic Acids Res.*, **36**, D637–40.
55. Harley, J.B., Alarcon-Riquelme, M.E., Criswell, L.A., et al. 2008, Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci, *Nat. Genet.*, **40**, 204–10.
56. Nath, S.K., Han, S., Kim-Howard, X., et al. 2008, A nonsynonymous functional variant in integrin-alpha(M) (encoded by ITGAM) is associated with systemic lupus erythematosus, *Nat. Genet.*, **40**, 152–4.
57. Kozyrev, S.V., Abelson, A.K., Wojcik, J., et al. 2008, Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus, *Nat. Genet.*, **40**, 211–6.
58. Goyette, P., Lefebvre, C., Ng, A., et al. 2008, Gene-centric association mapping of chromosome 3p implicates MST1 in IBD pathogenesis, *Mucosal Immunol.*, **1**, 131–8.
59. Johnson, A.D. and O'Donnell, C.J. 2009, An open access database of genome-wide association results, *BMC Med. Genet.*, **10**, 6.
60. Wellcome Trust Case Control Consortium 2007, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, *Nature*, **447**, 661–78.
61. Todd, J.A., Walker, N.M., Cooper, J.D., et al. 2007, Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes, *Nat. Genet.*, **39**, 857–64.
62. Plenge, R.M., Cotsapas, C., Davies, L., et al. 2007, Two independent alleles at 6q23 associated with risk of rheumatoid arthritis, *Nat. Genet.*, **39**, 1477–82.
63. Remmers, E.F., Plenge, R.M., Lee, A.T., et al. 2007, STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus, *N. Engl. J. Med.*, **357**, 977–86.
64. Graham, D.S., Graham, R.R., Manku, H., et al. 2008, Polymorphism at the TNF superfamily gene TNFSF4 confers susceptibility to systemic lupus erythematosus, *Nat. Genet.*, **40**, 83–9.
65. Ueda, H., Howson, J.M., Esposito, L., et al. 2003, Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease, *Nature*, **423**, 506–11.
66. Ikegami, H., Awata, T., Kawasaki, E., et al. 2006, The association of CTLA4 polymorphism with type 1 diabetes is concentrated in patients complicated with autoimmune thyroid disease: a multicenter collaborative study in Japan, *J. Clin. Endocrinol. Metab.*, **91**, 1087–92.

FEBS Letters 583 (2009) 1662–1667

ELSEVIER

FEBS Letters

journal homepage: www.FEBSLetters.org

Minireview

From transcriptome analysis to immunogenomics: Current status and future direction

Osamu Ohara *

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

ARTICLE INFO

Article history:
Received 16 February 2009
Revised 1 April 2009
Accepted 14 April 2009
Available online 18 April 2009

Edited by Miguel De la Rosa

Keywords:
Functional genomics
Immunogenomics
mRNA profiling
Protein profiling
Single-cell analysis

ABSTRACT

In 1994, we pioneered a complementary DNA (cDNA) sequencing project that aimed to predict the primary structures of unknown human proteins. Although our cDNA project was focused on the sequencing of large cDNAs, the following cDNA sequencing projects conducted by other groups have more extensively characterized mammalian transcriptome. In parallel, many groups have made a tremendous amount of effort to develop various resources for functional human genomics. In this context, to demonstrate the power of functional genomic approaches in practice, we have applied them for a comprehensive understanding of the immune system, which we term 'immunogenomics'. This mini-review first describes the historical background of our cDNA project and then provides perspectives on the present and future of immunogenomics based on our experiences.

1. Historical background: dawn of mammalian transcriptomic and proteomic analyses

In 1994, at Kazusa DNA Research Institute (KDRI), we initiated the analysis of the complete sequence of randomly sampled human complementary DNA (cDNAs) [1]. In this project, we intended to determine entire sequences of cDNAs and thereby to discover uncharacterized human protein-coding sequences encoded by the human genome; at that time, human cDNA analysis was mainly focused on the collection of expressed sequence tags [2]. The unique feature of our cDNA project is that the sequencing efforts have been focused on long cDNAs (>4 kb) that encode large proteins (>1000 amino acid residues) [3]. The novel human genes thus identified by our cDNA project were systematically designated as 'KIAA' plus a four-digit number; the number of 'KIAA' genes has today reached 2038 [4]. Subsequently, following our example, other research groups initiated highly comprehensive efforts to completely sequence mammalian (particularly human and mouse) cDNAs; although I cannot refer to each of the mammalian cDNA projects because of space limitation, many novel and unexpected aspects of the mammalian transcriptome have been revealed by their projects [5]. Together with the genomic sequence data, the accumulated data from such comprehensive analyses of mammalian cDNAs continue to serve as a solid basis for deciphering the mammalian transcriptome and proteome.

With the availability of the draft of the human genome sequence in 2001, cDNA analysis for discovering new human genes inevitably entered an end-game phase, at least for protein-coding genes. Consequently, we decided to terminate the first phase of the human cDNA project and to advance beyond the identification of transcribed sequences. From our perspective, it was natural to focus on the functional characterization of KIAA genes and their products in the second phase of our cDNA project because more than half of KIAA genes were not functionally annotated although we expected them to have certain special functions in mammals. Interestingly enough, large proteins encoded by large genes such as KIAA genes are expected to play a special and crucial role in the mammalian system [3]. This shift in focus from the transcriptome to the proteome was also a trend in the research community at that time. However, our practical concern was how to take advantage of the accumulated cDNA information and clone resources to reach our ultimate goal—a comprehensive understanding of the structure and function of the biological system. In

Abbreviations: KDRI, Kazusa DNA Research Institute; PID, primary immunodeficiency disease; cDNA, complementary DNA; 2DE, two-dimensional electrophoresis; RAPID, Resource of Asian Primary Immunodeficiency Diseases; RefDIC, Reference genomics Database of Immune Cells; ORF, protein-coding region; ORFeome, a complete set of protein-coding regions on the genome

* Address: Department of Human Genome Research, Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan. Fax: +81 438 52 3914.

E-mail addresses: ohara@kazusa.or.jp, oosamu@rcai.riken.jp

0014-5793/$36.00 © 2009 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2009.04.021

other words, we needed to design an approach to analyze gene function at the protein level via the analysis of cDNA because the true functional players in the biological system are proteins. In this regard, we believed that the cDNA clones thus accumulated in our project served as an indispensable and versatile resource for the functional genomics research community. After considerable deliberations, we decided to take an approach called 'affinity proteomics' [6]. In this approach, a large set of recombinant proteins derived from genes of interest are generated using their cDNA clones and used as antigens to raise antibodies. The raised antibodies are then used for characterization of the gene products in vivo by immunohistochemistry, protein blotting analysis, antibody array assays, pull-down experiments of the protein complex, and so on. We conducted such a project from 2002 to 2006, raising approximately 2000 antibodies against mouse KIAA proteins and analyzing these proteins in vivo by using these antibodies [7]. A similar, but more extensive and sophisticated, affinity proteomics approach is now being arranged for human gene products through an international collaboration of other groups [8].

The numerous completely sequenced cDNA clones are now indispensable reagents for experimental exploration in mammalian genomics. To support this impetus in research, a large number of protein-coding regions (which are conventionally designated as ORFs, although ORF was originally an abbreviation for open reading frame) excised from whole cDNAs have been cloned into vectors compatible with high-throughput cloning systems in the 'ORFeome project' [9,10]. The objectives of this project are exactly those envisioned by most human cDNA sequencing projects and are thus being realized through the means of international collaborations (Ref. [11], http://www.orfeomecollaboration.org/). While ORF clones thus shared by international collaboration are based on a recombination cloning method, it is worth noting that our ORF clones are prepared by a restriction enzyme-based high-throughput cloning system (Flexi Cloning System, Promega Co., WI, USA) and the HaloTag protein-tagging technology (Table 1; http://www.kazusa.or.jp/kop/dsearch-e/) [12].

In this way, dramatic progress has been made in mammalian transcriptomics and proteomics in the last decade. As an example, the outcomes of the projects conducted by KDRI are summarized in Table 1. In the context of the extensive functional genomics resources and know-how at KDRI, we considered that it is our responsibility to demonstrate the power of functional genomics in real-world biology. While there are many biological fields that could benefit from fully exploiting genomic resources, we have decided to become involved in immunological research. Therefore, this minireview hereafter describes the aim of our functional genomics research in immunology, termed 'immunogenomics', along with a perspective of the current status and future direction of our immunogenomics research.

2. What does immunogenomics comprise?

Genomics is a relatively new term in biology and encompasses a broad range of biological studies. Because a genome is the sum total of all the genes in an individual organism, genomics has been regarded as a high-throughput and quantitative form of biology. The advent of genomics has spawned the neologism 'omics' because the characteristics of genomics are quite distinct from those of conventional biology. According to the Omics.org portal site (http://omics.org/), 'omics' takes a certain philosophical perspective, which suggests a trend toward an integrative, rather than a reductionist, approach in biology. Interestingly, during the early emerging phase of the use of the term 'omics', some researchers emphasized the new aspects of a hypothesis-free (or data-driven) approach in biology with some others negating it [13–16]. As pointed out in [15], however, hypothesis-driven and data-driven approaches are never mutually exclusive. An important point is that genomics produces a huge amount of data which enables us to begin our analysis using data, and not entirely using theory or hypothesis. In this sense, 'omics' has just provided us with a new gateway into complex biological events that are too complicated for hypothesizing in advance. In this regard, it is crucial that the genomics data are shared with the research community because the more the data the better it is across all genomic approaches. Without data sharing in the research community, the amounts of data available for individual research groups would be too small to have a data-driven approach.

Although used in a similar context as that of 'omics', the term 'systems biology' is frequently used in a more detailed and specialized context than the term 'omics'. For example, the definition of 'systems biology' in Wikipedia (http://en.wikipedia.org/wiki/Systems_biology) is as follows: 'A relatively new biological study field that focuses on the systematic study of complex interactions in biological systems, thus using a new perspective (integration instead of reduction) to study them'. Thereafter, it has been stated that one of the goals of systems biology is to discover new emergent properties that may arise from the systemic view used by this discipline in order to better understand the entirety of processes that happen in a biological system. The phrase 'to better understand the entirety' is considerably popular among biologists, thus increasing the popularity of 'systems biology'. However, this simply reminds me of an old book entitled 'Cybernetics' [17], which I was very excited about more than 30 years ago. The main difference between the era of 'systems biology' and the era described in Cybernetics is that the genome information provides us with an almost-complete list of functional elements in the biological system, i.e. the proteins. In this respect, a recent study by Kitano's group might be regarded as an orthodox systems-biological descendent of Cybernetics [18]. To further move beyond 'Cybernetics' and better understand the entirety of the biological system from the complete list of proteins, we may need to devise a novel and logical framework; this is an exciting challenge in the post-genomic era.

In this minireview, 'immunogenomics' is used to include genomics focused on immunology, although the usage of this term has been fairly recent and infrequent in literature. Because the immune system in mammals is one of the most complex biological systems, we consider it the most challenging target field to demonstrate the power of genomic approaches. Furthermore, we believe

Table 1
Databases generated by the Kazusa mammalian cDNA projects.

Database name	URL	Description
HUGE	http://www.kazusa.or.jp/huge/	Human unidentified-gene encoded large protein database
ROUGE	http://www.kazusa.or.jp/rouge/	Roudent unidentified-gene encoded large protein database
NEDO	http://www.kazusa.or.jp/NEDO/	Database for human long cDNAs in spleen
InGaP	http://webcreate.kazusa.or.jp/create/	Database for gene expression/protein profiles obtained by using mouse KIAA antibodies
InCeP	http://webcreate.kazusa.or.jp/create/	Database for intracellular pathways based on mouse KIAA protein-protein interactions
KOP	http://www.kazusa.or.jp/kop/	Database for Kazusa ORF clones and expression clones for the ORF-HaloTag® fusions[a]

[a] http://www.promega.com/applications/immdet/imaging/halotag/.

that the application of genomics in order to solve difficult problems might bring about a breakthrough in immunology because it will effectively complement conventional immunological approaches. Although there are various types of important properties of the immune system to be explored via genomic approaches, we have hereafter focused our discussion on mRNA and/or protein profiling of the immune cells because of space constraints.

3. Immunogenomics resources and their application to real-world immunology

3.1. Resources for immunogenomics

As described above, a data-driven approach can be a good alternative to a conventional hypothesis-driven one, particularly when the phenomenon of interest is too complex to establish hypotheses in advance. However, the data-driven approach cannot be realized without high-quality genomic data. In this respect, transcriptomic data are the most practically suited to being comprehensively tackled with current technology. For this purpose, DNA microarray technology has already opened a way to accumulate large amounts of data regarding genome-wide gene expression profiles. The description of biological systems with tens of thousands of different mRNA levels is highly sensitive to the state of those systems. Although mRNA profiles obtained by DNA microarray analyses are not absolutely quantitative and somewhat depend upon the types of DNA microarray platforms, DNA microarray technology has already been demonstrated to uncover intriguing features of gene expression profiles in various biological events. However, when we planned to implement a genomics basis in immunology, there was no open-access database of mRNA profiles of immune cells. Thus, we first worked on generating a reference database of gene expression in immune cells. It was our plan to freely provide the research community with the resulting informational resource of gene expression of immune cells because data sharing plays a pivotal role in genomics. To achieve this purpose, we invited many researchers in the RIKEN Research Centre for Allergy and Immunology (http://www.rcai.riken.jp) to collaborate with us.

As we began to construct this immunogenomics database, we were concerned that mRNA profiles alone would not be sufficient to reveal the molecular mechanisms underlying the observed biological events. This is mainly because mRNA levels are not always directly relevant to biological phenomena; mRNA is a template for protein synthesis and is not itself a functional element in most cases. Protein profiles have a closer relevance to biological events than do mRNA profiles because it is the proteins that directly govern biological events. The scarcity of quantitative proteomic data was thus a serious concern when analyzing the immune system from an 'omics' viewpoint. To address this problem practically, we previously generated quantitative proteomic maps based on two-dimensional gel electrophoresis (2DE) and included these quantitative data in our immunogenomics database [19].

In order to integrate the 'omics' data obtained for various immune cells, we annotated each sample with a controlled vocabulary and implemented a relational database that included the quantitative mRNA and protein profiling data as described above. Our immunogenomics database, termed 'Reference genomics Database of Immune Cells' (RefDIC), offers a web-based query interface and user-friendly facilities for visualizing data, and allows all of the raw data to be downloaded [20]. 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 [20]. The number of mRNA/protein profiles in RefDIC continues to grow and its data-viewing functions have also been enhanced continuously (Fig. 1).

On similar lines, several groups have also constructed their own gene expression databases to date [21–23]. Among them, the Immunological Genome project (the ImmGen project, http://www.immgen.org) aims to construct an open-access database focusing on immune cells; their data are also shared with the research community, like ours [24]. Interestingly enough, on the basis of their data as well as ours, two-thirds of the genome has been found to be active in at least one immune cell type; moreover, only less than 1% of genes are expressed exclusively in a given type of immune cell. This implies that most immunological phenomena should be interpreted in the framework of the whole system rather than as functions of a small set of genes specific to a given type of immune cells. In other words, characteristic phenotypes of immune cells are likely to result mainly from characteristics of the whole system of the gene expression network.

3.2. Technological problems in proteome analyses in immunogenomics

Although protein profile data based on 2DE have been included, an obvious weakness of the RefDIC datasets is the low comprehensiveness of the proteome data. Our quantitative profiles in RefDIC based on 2DE comprise the profile of only several hundred proteins that occur in abundance in immune cells. This is generally very problematic in practice and needs to be resolved. Advances in mass spectrometry-based proteomics have made it possible to obtain absolute quantities of several thousand proteins [25], but these techniques are still in their infancy from the viewpoint of comprehensive protein profiling. In addition, many functionally important proteins in immunology are not abundant, or even very rare, in many cases. Thus, it would be difficult to comprehensively analyse these functionally important proteins (e.g. cytokines/chemokines and transcription factors) by even the most advanced mass spectrometry-based proteomic approaches. One possible solution for this is to again take an affinity proteomics approach as described in Section 1. Although it is still impossible to comprehensively monitor protein profiles in a genome-wide fashion, current antibody array technology enables us to monitor more than several tens of protein levels simultaneously with very high sensitivity. We are now collecting cytokine profiles in this manner and will soon be depositing these data in RefDIC. Similarly, we will be able to monitor the protein levels of transcription factors that are relevant to immunological events. Another solution to this issue is to monitor the levels of mRNA engaged in translation, instead of the total mRNA present in immune cells. In practice, this can be achieved by mRNA profiling analysis of polysomal RNAs. Through this approach, we can monitor the mRNA translation profiles that are expected to correlate well with the rates of protein production. Thus, although this approach cannot monitor the levels of accumulated protein in immune cells, the mRNA profiles thus obtained are easier to interpret in terms of protein production. Taking this approach, we could partially identify post-transcriptional gene regulation in a macrophage-like cell line elicited by a lipopolysaccharide [26].

A technological concern intrinsic to both transcriptomic and proteomic data is that they are obtained only as snapshots and as averages of an immune cell population. As demonstrated recently [27], bio-imaging may provide a partial answer to this concern because quantitative bio-imaging data are obtained for single cells, not populations, and can be followed over time. Thus, we need a new logic and framework to deal with these new types of data when integrating single-cell based data with conventional genomic data. This raises a difficult but exciting technological challenge for the future, and will be discussed in detail in Section 4.

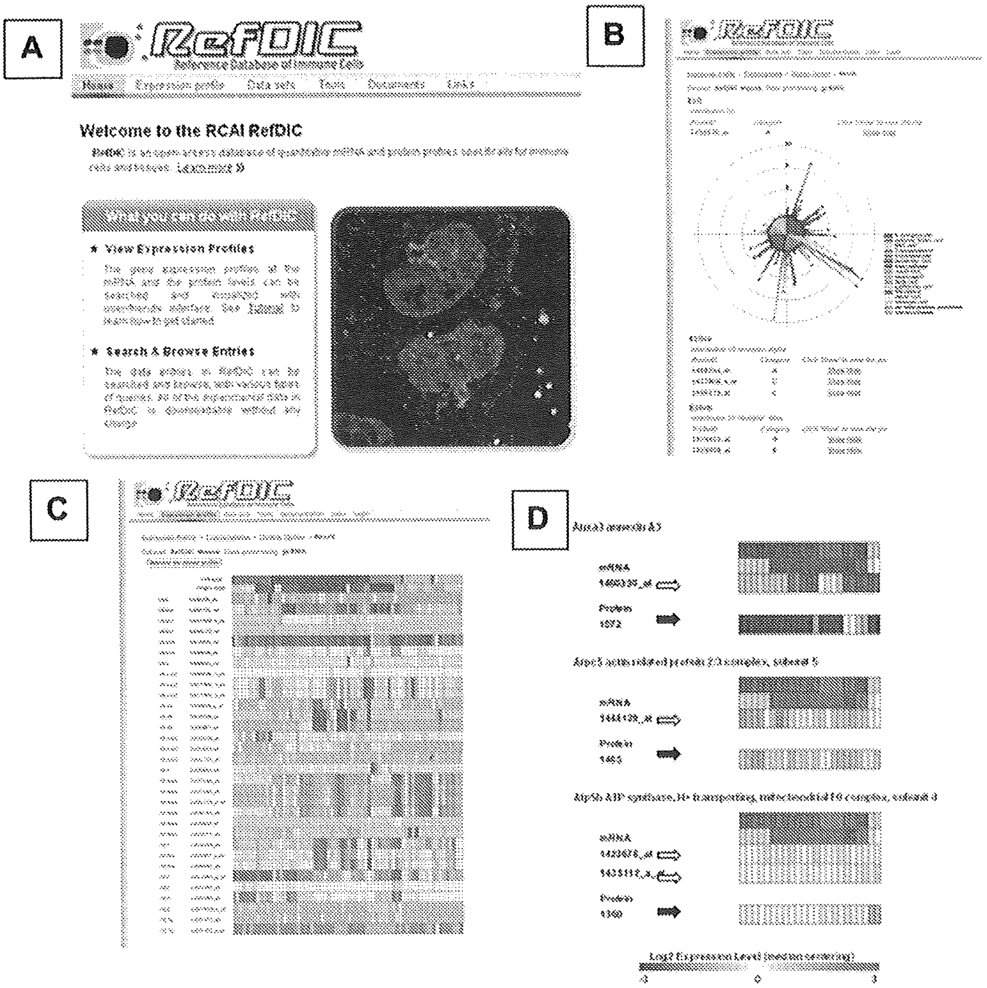
A
B
C
D
RefDIC
Welcome to the RCAI RefDIC
What you can do with RefDIC
View Expression Profiles
Search & Browse Entries
mRNA
Protein
Log2 Expression Level (median centering)
-3
0
3

Fig. 1. Snapshots of RefDIC. Panel A shows a top page of RefDIC (http://refdic.rcai.riken.jp/). Panels B and C display mRNA profiles in a radar chart format and a heat-map format, respectively. Panel D is an example of comparison of protein and mRNA profiles (indicated by filled and open arrows, respectively) in various samples in a heat-map format.

3.3. Application of immunogenomics for analysis of immunological diseases

While there are some technological limitations as described above, the use of the accumulated mRNA profile data has already proven to be very powerful in practice. One of the most exciting cases is the identification of four genes which convert fibroblasts to induced pluripotent stem cells [28]. This is a case of success through 'signature' identification. A 'signature' is a set of genes identified by comparison of mRNA profiles of cells under conditions defined from a perceived biological logic as described in [22]. In many cases, signature analysis is very informative in immunology and is frequently performed in a hypothesis-free manner. In addition to signature analysis, integration of multiple types of genomic data is also considered to be highly useful [29]. However, from my perspective, it is more intriguing to apply immunogenomics to highly complex immune events which are beyond the reach of current reductive approaches. In this regard, disease studies can be a touchstone since they require highly integrative approaches to fill the gap between genotypes and phenotypes. It would be important to know how far we can assess the relationship between genotypes and phenotypes via genomic approaches since disease studies are definitely multi-scale analyses and require highly integrative knowledge of biological systems. To achieve this, we first need a solid basis of disease-related information. As an example of disease studies, we have recently begun constructing a database of primary immunodeficiency diseases (PIDs). Because all the genetic causes of PIDs identified so far are single gene mutations, the information required is more straightforward than in other immunological diseases. Thus, we have constructed a web-based compendium of molecular alterations in PID, named the Resource of Asian Primary Immunodeficiency Diseases (RAPID), which is available as a worldwide web resource at http://rapid.rcai.riken.jp/ [30]. RAPID hosts information on manually curated sequence variations, gene expression profiles, immune pathway information, and various lines of information at the protein level of PID genes (Fig. 2). RAPID is in its early infancy, but will certainly make it easier to integrate various lines of information and data mining for the analysis of PID. However, most of the information currently stored in RAPID is at the molecular level. Thus, we must obtain more information regarding the phenotypes of PIDs in relation with genotypes in the future. Only then will we be able to tackle the challenging issue of understanding the entirety better by fully exploiting immunogenomics approaches.

4. Future direction: to develop single-cell based genomics for a comprehensive systematic understanding of the immune system

As described earlier, the amounts of information accumulated in immunogenomics have continuously increased and are already

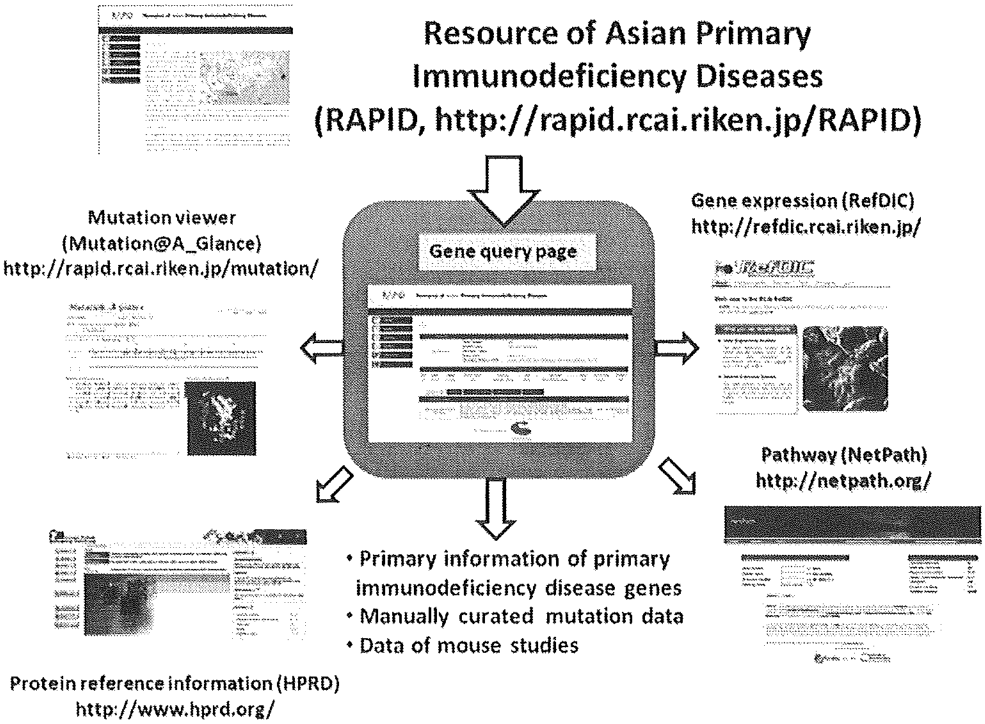
Resource of Asian Primary Immunodeficiency Diseases
(RAPID, http://rapid.rcai.riken.jp/RAPID)
Gene query page
Mutation viewer
(Mutation@A_Glance)
http://rapid.rcai.riken.jp/mutation/
Gene expression (RefDIC)
http://refdic.rcai.riken.jp/
Pathway (NetPath)
http://netpath.org/
• Primary information of primary immunodeficiency disease genes
• Manually curated mutation data
• Data of mouse studies
Protein reference information (HPRD)
http://www.hprd.org/

Fig. 2. Integration of various genomic information for PID genes on RAPID. A screen shot of the top page of RAPID is shown (http://rapid.rcai.riken.jp/RAPID/) at the top of this figure. After selecting a PID gene from the gene query section, various types of information for the PID gene are linked. Snapshots of representative linked sites are shown with their URLs.

being utilized in real-world immunology. The power of the genomic approach is now widely accepted in general, but the limitations and pitfalls of current genomics technologies, particularly for the analysis of biological systems consisting of multiple cells such as the immune system, have also become evident. Among them, how to characterize multi-scale biological systems with heterogeneous cell populations at the molecular level is one of the most critical problems in functional genomics. From my perspective, we will confront this difficult problem most specifically in immunology because dynamic changes in cell population and interactions among heterogeneous cells play an essential role in immunological events.

The limitations and pitfalls described above originate from current technological limitations of molecular profiling. For example, although cell designation by using a fluorescence-activated cell sorter has been used for many years and is believed to be robust, our experience as well as the findings of the early confirmation studies in the ImmGen project have indicated that the mRNA profile data obtained using identically designated cell populations in different laboratories vary considerably. Because we obtained the mRNA profiles as averages of cell populations, we cannot identify whether the inter-laboratory variability originates from heterogeneity in the collected cells or from intrinsic differences in individual cells. Thus, how to move beyond 'the myth of the average cell' is closely related to an essential problem, i.e. how to experimentally obtain a systems view of the immune system as discussed in [31]. More importantly, even if a cell population is composed of clonal cells, cell-to-cell variations do exist. It is interesting to note that phenotypic cell-to-cell variability within clonal populations could arise from the slow fluctuation of protein levels in mammalian cells [32,33]. Therefore, not only heterogeneity of cell types but also cell-to-cell variations of clonal cells must be taken into consideration to understand the immune system from a systems-biological viewpoint. This will also raise a new theoretical/philosophical problem of how to deal with biological systems consisting of heterogeneous cells, where a reductive approach seems almost useless.

From a technological viewpoint, a solution to this problem has already come from bio-imaging technology. The current imaging technology enables us to monitor the behavior of even a single molecule in a living cell in a quantitative manner. In addition, in vivo imaging technology has advanced rapidly and thereby we may soon be able to follow the behavior of single cells in vivo in real time. Nonetheless, as discussed by Albeck et al. [34], the most productive and reliable approach in practice is to integrate and compare data obtained by both population-based (or bulk) and single-cell approaches because at present, genomic data can be obtained only for a population and the data from single-cell approaches are full of noise. It would be interesting to observe what could happen if the simultaneous measurement of a large number of mRNA and/or protein levels in many single cells is possible. Single-cell mRNA/protein profiling will produce drastically increased amounts of quantitative data and enable us to characterize complex cell ensembles at the molecular level. Although Mar et al. pioneered this direction of analysis using a mesoscopic approach [35], more straightforward technologies for biochemical/genomic analyses of single cells are right around the corner [36,37]. Therefore, it is time to seriously consider the logic required to deal with these data with theoreticians and/or computational scientists. Together with the accumulated genomic data, these new genomic data at the single-cell level will deepen our knowledge of multicellular biological systems, although the research community might be thrown into confusion at the beginning. This situation might bear some analogy to the early phase of the emergence of statistical mechanics from thermodynamics. In the next 10 years, biology may change drastically in this direction. In my outlook, immunogenomics will surely play a pivotal role in this process.

Acknowledgments

The author would like to express his gratitude to his numerous colleagues. In particular, Dr. Takahiro Nagase, Dr. Hisashi Koga, Dr. Reiko F. Kikuno, Dr. Hiroshi Kitamura, Dr. Yayoi Kimura and Mr. Atsushi Hijikata are sincerely acknowledged. My researches have been supported by grants for Kazusa DNA Research Institute and for RIKEN Research Center for Allergy and Immunology. This work was also supported in part by the Global COE Program (Global Centre for Education and Research in Immune System Regulation and Treatment), the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. As for immunogenomics for PIDs, the author would like to thank Dr. Akhilesh Pandey, Dr. Sujatha Mohan and their colleagues for their great contributions to this project. The construction of the RAPID database was supported in part by Special coordination Funds for Promoting Science and Technology by MEXT.

References

[1] Nomura, N., Miyajima, N., Sazuka, T., Tanaka, A., Kawarabayasi, Y., Sato, S., Nagase, T., Seki, N., Ishikawa, K.-I. 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.

[2] Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C. and Venter, J.C. (1992) Sequence identification of 2375 human brain genes. Nature 355, 632–634.

[3] Ohara, O., Nagase, T., Ishikawa, K., Nakajima, D., Ohira, M., Seki, N. and Nomura, N. (1997) Construction and characterization of human brain cDNA libraries suitable for analysis of cDNA clones encoding relatively large proteins. DNA Res. 4, 53–59.

[4] Nagase, T., Koga, H. and Ohara, O. (2006) Kazusa mammalian cDNA resources: towards functional characterization of KIAA gene products. Brief Funct. Genomic Proteomic 5, 4–7.

[5] Carninci, P., Yasuda, J. and Hayashizaki, Y. (2008) Multifaceted mammalian transcriptome. Curr. Opin. Cell Biol. 20, 274–280.

[6] Uhlen, M. (2008) Affinity as a tool in life science. Biotechniques 44, 649–654.

[7] Koga, H., Yuasa, S., Nagase, T., Shimada, K., Nagano, M., Imai, K., Ohara, R., Nakajima, D., Murakami, M., Kawai, M., Miki, F., Magae, J., Inamoto, S., Okazaki, N. and Ohara, O. (2004) A comprehensive approach for establishment of the platform to analyze functions of KIAA proteins II: public release of inaugural version of InGaP database containing gene/protein expression profiles for 127 mouse KIAA genes/proteins. DNA Res. 11, 293–304.

[8] Taussig, M.J., Stoevesandt, O., Borrebaeck, C.A., Bradbury, A.R., Cahill, D., Cambillau, C., de Daruvar, A., Dubel, S., Eichler, J., Frank, R., Gibson, T.J., Gloriam, D., Gold, L., Herberg, F.W., Hermjakob, H., Hoheisel, J.D., Joos, T.O., Kallioniemi, O., Koegl, M., Konthur, Z., Korn, B., Kremmer, E., Krobitsch, S., Landegren, U., van der Maarel, S., McCafferty, J., Muyldermans, S., Nygren, P.A., Palcy, S., Pluckthun, A., Polic, B., Przybylski, M., Saviranta, P., Sawyer, A., Sherman, D.J., Skerra, A., Templin, M., Ueffing, M. and Uhlen, M. (2007) Proteome binders: planning a European resource of affinity reagents for analysis of the human proteome. Nat. Meth. 4, 13–17.

[9] Goshima, N., Kawamura, Y., Fukumoto, A., Miura, A., Honma, R., Satoh, R., Wakamatsu, A., Yamamoto, J., Kimura, K., Nishikawa, T., Andoh, T., Iida, Y., Ishikawa, K., Ito, E., Kagawa, N., Kaminaga, C., Kanehori, K., Kawakami, B., Kenmochi, K., Kimura, R., Kobayashi, M., Kuroita, T., Kuwayama, H., Maruyama, Y., Matsuo, K., Minami, K., Mitsubori, M., Mori, M., Morishita, R., Murase, A., Nishikawa, A., Nishikawa, S., Okamoto, T., Sakagami, N., Sakamoto, Y., Sasaki, Y., Seki, T., Sono, S., Sugiyama, A., Sumiya, T., Takayama, T., Takayama, Y., Takeda, H., Togashi, T., Yahata, K., Yamada, H., Yanagisawa, Y., Endo, Y., Imamoto, F., Kisu, Y., Tanaka, S., Isogai, T., Imai, J., Watanabe, S. and Nomura, N. (2008) Human protein factory for converting the transcriptome into an in vitro-expressed proteome. Nat. Meth. 5, 1011–1017.

[10] Rual, J.F., Hirozane-Kishikawa, T., Hao, T., Bertin, N., Li, S., Dricot, A., Li, N., Rosenberg, J., Lamesch, P., Vidalain, P.O., Clingingsmith, T.R., Hartley, J.L., Esposito, D., Cheo, D., Moore, T., Simmons, B., Sequerra, R., Bosak, S., Doucette-Stamm, L., Le Peuch, C., Vandenhaute, J., Cusick, M.E., Albala, J.S., Hill, D.E. and Vidal, M. (2004) Human ORFeome version 1.1: a platform for reverse proteomics. Genome Res. 14, 2128–2135.

[11] Temple, G., Lamesch, P., Milstein, S., Hill, D.E., Wagner, L., Moore, T. and Vidal, M. (2006) From genome to proteome: developing expression clone resources for the human genome. Human Mol. Genet. 15 (1), R31–R43.

[12] Nagase, T., Yamakawa, H., Tadokoro, S., Nakajima, D., Inoue, S., Yamaguchi, K., Itokawa, Y., Kikuno, R.F., Koga, H. and Ohara, O. (2008) Exploration of human ORFeome: high-throughput preparation of ORF clones and efficient characterization of their protein products. DNA Res. 15, 137–149.

[13] Goodman, L. (1999) Hypothesis-limited research. Genome Res. 9, 673–674.

[14] Kell, D.B. and Oliver, S.G. (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. Bioessays 26, 99–105.

[15] Lastowski, K. and Makalowski, W. (2000) Methodological function of hypotheses in science: old ideas in new cloth. Genome Res. 10, 273–274.

[16] Selinger, D.W., Wright, M.A. and Church, G.M. (2003) On the complete determination of biological systems. Trends Biotechnol. 21, 251–254.

[17] Wiener, N. (1961) Cybernetics, 2nd ed, The MIT Press, Cambridge, Massachusetts.

[18] Oda, K. and Kitano, H. (2006) A comprehensive map of the toll-like receptor signaling network. Mol. Syst. Biol. 2, 0015.

[19] Kimura, Y., Yokoyama, R., Ishizu, Y., Nishigaki, T., Murahashi, Y., Hijikata, A., Kitamura, H. and Ohara, O. (2006) Construction of quantitative proteome reference maps of mouse spleen and lymph node based on two-dimensional gel electrophoresis. Proteomics 6, 3833–3844.

[20] Hijikata, A., Kitamura, H., Kimura, Y., Yokoyama, R., Aiba, Y., Bao, Y., Fujita, S., Hase, K., Hori, S., Ishii, Y., Kanagawa, O., Kawamoto, H., Kawano, K., Koseki, H., Kubo, M., Kurita-Miki, A., Kurosaki, T., Masuda, K., Nakata, M., Oboki, K., Ohno, H., Okamoto, M., Okayama, Y.J.O.W., Saito, H., Saito, T., Sakuma, M., Sato, K., Seino, K., Setoguchi, R., Tamura, Y., Tanaka, M., Taniguchi, M., Taniuchi, I., Teng, A., Watanabe, T., Watarai, H., Yamasaki, S. and Ohara, O. (2007) Construction of an open-access database that integrates cross-reference information from the transcriptome and proteome of immune cells. Bioinformatics 23, 2934–2941.

[21] Abbas, A.R., Baldwin, D., Ma, Y., Ouyang, W., Gurney, A., Martin, F., Fong, S., van Lookeren Campagne, M., Godowski, P., Williams, P.M., Chan, A.C. and Clark, H.F. (2005) Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. Genes Immun. 6, 319–331.

[22] Hyatt, G., Melamed, R., Park, R., Seguritan, R., Laplace, C., Poirot, L., Zucchelli, S., Obst, R., Matos, M., Venanzi, E., Goldrath, A., Nguyen, L., Luckey, J., Yamagata, T., Herman, A., Jacobs, J., Mathis, D. and Benoist, C. (2006) Gene expression microarrays: glimpses of the immunological genome. Nat. Immunol. 7, 686–691.

[23] Splendiani, A., Brandizi, M., Even, G., Beretta, O., Pavelka, N., Pelizzola, M., Mayhaus, M., Foti, M., Mauri, G. and Ricciardi-Castagnoli, P. (2007) The genopolis microarray database. BMC Bioinform. 8 (Suppl. 1), S21.

[24] Heng, T.S. and Painter, M.W.Immunological Genome Project Consortium (2008) The Immunological Genome Project: networks of gene expression in immune cells. Nat. Immunol. 9, 1091–1094.

[25] Cox, J. and Mann, M. (2007) Is proteomics the new genomics? Cell 130, 395–398.

[26] Kitamura, H., Ito, M., Yuasa, T., Kikuguchi, C., Hijikata, A., Takayama, M., Kimura, Y., Yokoyama, R., Kaji, T. and Ohara, O. (2008) Genome-wide identification and characterization of transcripts translationally regulated by bacterial lipopolysaccharide in macrophage-like J774.1 cells. Physiol. Genomics 33, 121–132.

[27] Cohen, A.A., Geva-Zatorsky, N., Eden, E., Frenkel-Morgenstern, M., Issaeva, I., Sigal, A., Milo, R., Cohen-Saidon, C., Liron, Y., Kam, Z., Cohen, L., Danon, T., Perzov, N. and Alon, U. (2008) Dynamic proteomics of individual cancer cells in response to a drug. Science 322, 1511–1516.

[28] Takahashi, K. and Yamanaka, S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126, 663–676.

[29] Quackenbush, J. (2007) Extracting biology from high-dimensional biological data. J. Exp. Biol. 210, 1507–1517.

[30] Keerthikumar, S., Raju, R., Kandasamy, K., Hijikata, A., Ramabadran, S., Balakrishnan, L., Ahmed, M., Rani, S., Selvan, L.D., Somanathan, D.S., Ray, S., Bhattacharjee, M., Gollapudi, S., Ramachandra, Y.L., Bhadra, S., Bhattacharyya, C., Imai, K., Nonoyama, S., Kanegane, H., Miyawaki, T., Pandey, A., Ohara, O. and Mohan, S. (2009) RAPID: Resource of Asian Primary Immunodeficiency Diseases. Nucl. Acids Res. 37, D863–D867.

[31] Levsky, J.M. and Singer, R.H. (2003) Gene expression and the myth of the average cell. Trends Cell Biol. 13, 4–6.

[32] Chang, H.H., Hemberg, M., Barahona, M., Ingber, D.E. and Huang, S. (2008) Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature 453, 544–547.

[33] Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Liron, Y., Rosenfeld, N., Danon, T., Perzov, N. and Alon, U. (2006) Variability and memory of protein levels in human cells. Nature 444, 643–646.

[34] Albeck, J.G., MacBeath, G., White, F.M., Sorger, P.K., Lauffenburger, D.A. and Gaudet, S. (2006) Collecting and organizing systematic sets of protein data. Nat. Rev. Mol. Cell Biol. 7, 803–812.

[35] Mar, J.C., Rubio, R. and Quackenbush, J. (2006) Inferring steady state single-cell gene expression distributions from analysis of mesoscopic samples. Genome Biol. 7, R119.

[36] Sasuga, Y., Iwasawa, T., Terada, K., Oe, Y., Sorimachi, H., Ohara, O. and Harada, Y. (2008) Single-cell chemical lysis method for analyses of intracellular molecules using an array of picoliter-scale microwells. Anal. Chem. 80, 9141–9149.

[37] Svahn, H.A. and van den Berg, A. (2007) Single cells or large populations? Lab Chip 7, 544–546.

ELSEVIER

Experimental Hematology

Experimental Hematology 2009;37:701–714

Multistep pathogenesis of leukemia via the MLL-AF4 chimeric gene/Flt3 gene tyrosine kinase domain (TKD) mutation-related enhancement of S100A6 expression

Hiroki Yamaguchi[a], Hideki Hanawa[b], Naoya Uchida[a,b], Mitsuharu Inamai[a], Kazuhiro Sawaguchi[a], Yoshio Mitamura[a], Takashi Shimada[b], Kazuo Dan[a], and Koiti Inokuchi[a]

[a]Division of Hematology, Department of Internal Medicine, Nippon Medical School, Tokyo, Japan; [b]Department of Biochemistry and Molecular Biology, Nippon Medical School, Tokyo, Japan

(Received 28 October 2008; revised 2 February 2009; accepted 9 February 2009)

Objective. **Concerning MLL-AF4 leukemogenesis, previous mouse models suggest that the tumorigenesis capacity of MLL-AF4 alone is insufficient for causing leukemia. Based on the finding that an Fms-like tyrosine kinase 3 (Flt3) gene mutation in the tyrosine kinase domain (TKD) was observed in approximately 15% of mixed lineage leukemia (MLL), we investigated synergistic leukemogenesis effects of the two genes in vitro.**

Materials and Methods. **In a mouse interleukin-3 (IL-3)−dependent cell line, 32Dc, expression of MLL-AF4 and mutant Flt3 was induced using a lentiviral vector. We analyzed apoptosis induction in the absence of IL-3 and the granulocyte colony-stimulating factor − related induction of differentiation, gene expression profiling, and the mechanism involved in the synergistic effects of MLL-AF4 and Flt3-TKD.**

Results. **Neither Flt3-expressing 32Dc ($32Dc^{Flt3\text{-}TKD}$) nor MLL-AF4−expressing 32Dc ($32Dc^{MLL\text{-}AF4}$) acquired IL-3−independent proliferative capacity in semisolid/liquid media. However, Flt3-TKD + MLL-AF4−expressing 32Dc ($32Dc^{Flt3\text{-}TKD+MLL\text{-}AF4}$) acquired a non − IL-3−dependent proliferative capacity by inhibiting apoptosis in the two media. The $32Dc^{Flt3\text{-}TKD}$ and $32Dc^{MLL\text{-}AF4}$ cells differentiated into granulocytes in the presence of granulocyte colony-stimulating factor. However, in the $32Dc^{Flt3\text{-}TKD+MLL\text{-}AF4}$ cells, there was no differentiation. Subsequently, we performed gene expression profiling. The enhancement of Hox genes expression was not identified. However, expression of S100A6 was synergistically enhanced in the presence of both MLL-AF4 and Flt3-TKD genes. Moreover, anti-S100A6 small interfering RNA downregulated leukemic proliferation.**

Conclusion. **We conclude that their synergistic enhancement of S100A6 expression plays an important role in MLL-AF4−associated leukemogenesis.**

Translocation of the long-arm q23 region (11q23) of chromosome number 11 is frequent in hematopoietic malignancies, and is detected in approximately 5% of patients with acute leukemia. In particular, this chromosomal abnormality is frequently observed in patients with infantile or secondary leukemia [1]. The mixed lineage leukemia (MLL) gene, which translocated from an area adjacent to the cleavage site of the translocation to the 11q23 region, was identified. MLL fusion genes resulting from translocation-related rearrangement play an important role in the pathogenesis of leukemia [2,3]. Previous studies have reported that the MLL gene fused with approximately 40 different partner genes. In particular, the translocation of t(4;11)(q21;q23) was frequent in infants and children with pro − B acute lymphoblastic leukemia (ALL) or pro-B/myeloid MLL [4]. In clinical practice, t(4;11)(q21;q23) is regarded as a poor prognostic factor [4,5].

With respect to the mechanism by which MLL-AF4 chimeric protein formed via t(4;11)(q21;q23) translocation causes leukemia, the following findings have been obtained: first, MLL-AF4−expressing cells resisted apoptosis mediated by serum starvation and CD95 [6,7]; second, when MLL-AF4 was knocked down using small interfering

Offprint requests to: Hiroki Yamaguchi, M.D., Ph.D., Division of Hematology, Department of Internal Medicine, Nippon Medical School, 1-1-5 Sendagi, Bunkyo-Ku, Tokyo 113-8603, Japan; E-mail: y-hiroki@fd6.so-net.ne.jp

0301-472X/09 $–see front matter. Copyright © 2009 ISEH - Society for Hematology and Stem Cells. Published by Elsevier Inc.
doi: 10.1016/j.exphem.2009.02.007

RNA (siRNA) in an MLL-AF4–expressing cell line, expressions of Hoxa7, Hoxa9, and Meis1 genes decreased, leading to loss of independent proliferative capacity [8]; and third, MLL-AF4–expressing model mice developed B-cell lymphoma or myeloid leukemia, differing from the de novo condition, via long-term follow-up after the appearance of a myeloproliferative disorder-like phenotype [9,10]. Therefore, MLL-AF4 has a tumorgenesis capacity, but a second hit may be necessary to cause leukemia.

Recently, it has been speculated that the combination of a gene abnormality increasing proliferative activity (class I) and that inhibiting differentiation (class II) is essential for the onset of leukemia [11]. Previous studies reported that expression of the Fms-like tyrosine kinase 3 (Flt3) (a receptor-type tyrosine kinase) gene was high in patients with MLL translocation-type leukemia on DNA microarray analysis [12], and that an Flt3 mutation in the tyrosine kinase domain (TKD) was detected in approximately 15% of patients with de novo MLL translocation-type leukemia [13,14]. In addition, use of Flt3 inhibitors, such as PKC412 and CEP701, in MLL translocation-type leukemia cell lines and its clinical specimens markedly inhibited cell proliferation [13,15–17], suggesting that, as a second hit, a class I mutation, Flt3-TKD, is closely involved in the mechanism by which MLL-AF4 causes leukemia. Ono et al. [18] combined mouse MLL-SEPT6 or MLL-ENL models with Flt3 internal tandem duplication (ITD), and found that these combinations caused acute leukemia earlier via their synergistic effects compared to MLL chimeric proteins or Flt3-ITD mutation alone.

However, as described here, Flt3-TKD is more than Flt3-ITD in patients with MLL translocation-type leukemia. We performed the in vitro molecular biological analysis of MLL-AF4 and Flt3-TKD, which are frequent among patients with MLL translocation-type leukemia, to verify the multistep pathogenesis of leukemia associated with the two genes.

Materials and methods

Cell line, growth factor, and antibodies

We purchased mouse IL-3–dependent 32Dc (32Dc), RS411 (positive for MLL-AF4), and MV411 (positive for MLL-AF4 and Flt3-ITD) cells from the American Type Culture Collection (ATCC; Manassas, VA, USA). Mouse erythroleukemia cells, HL60, K562, and KML-1 cells were purchased from Cell Bank, RIKEN BioResource Center (Ibaragi, Japan). Recombinant murine IL-3 and granulocyte colony-stimulating factor (G-CSF) were supplied by Kirin Pharma Co., Ltd. (Tokyo, Japan). Concerning antibodies, we used Flt3/Flk-2 antibody (Santa Cruz Biotechnology, Santa Cruz, CA, USA) as an anti-Flt3 antibody, MLLT2 antibody (Orbigen, San Diego, CA, USA) as an anti-MLL antibody, phospho-Flt3 (Tyr591) antibody (Cell Signaling Technology, Danvers, MA, USA) as an anti–phospho-Flt3 antibody, calcyclin (H-55) (Santa Cruz Biotechnology) as anti-S100A6 antibody, signal transducers and activators of transcription 5 (STAT5) (9363) (Cell Signaling Technology) as anti-STAT5 antibody, phospho-STAT5 (Tyr694) (9351) (Cell Signaling Technology) as anti–phospho-STAT5 antibody, Pim-2 (1D12) (Santa Cruz Biotechnology) as anti – Pim-2 antibody, and glyceraldehyde-3-phosphate dehydrogenase (GAPDH)–loading control (Imgenex, San Diego, CA, USA) as an anti-GAPDH antibody.

MLL-AF4 cDNA cloning

We successfully established a cell line (TK cell) from pro – B ALL patients with t(4;11) (q21;q23) translocation. We prepared complementary DNA (cDNA) from the cell line, and amplified seven fragments using reverse transcription polymerase chain reaction (RT-PCR). After cloning was performed using TA cloning (Invitrogen, Carlsbad, CA, USA), full-length MLL-AF4 cDNA was prepared by ligating each fragment.

Plasmid construction

MLL-AF4 cDNA was inserted into a lentiviral vector, pCL20cCMp + EF1a-GFP, with a green fluorescent protein (GFP)–expressing marker at the C-terminal. In this experiment, we selected the lentiviral vector because insertion-related mutations around the starting point of oncogene transcription are less frequent than in retroviral vectors. As MLL-AF4 cDNA was large (7026 bp), we inserted a cytomegalovirus enhancer next to the murine stem cell virus promoter to improve expression efficiency (Fig. 1A). Flt3 cDNA was purchased from OriGene Inc. (Rockville, MD, USA). Using a site-direct mutagenesis kit (Invitrogen), we prepared Flt3-ITD/Flt3-TKD D835Y mutations. Each Flt3 cDNA was inserted into pCL20cCMp + EF1a-Ds-Red (Fig. 1A), in which the C-terminal was substituted for a Ds-Red–expressing marker, pDsRed-Express-1 (Clontech, Mountain View, CA, USA), to achieve coexpression with MLL-AF4. Concerning lentivirus preparation was described previously [19]. Lentiviruses expressing MLL-AF4 and each Flt3 cDNA, which we prepared, were transduced to 32Dc cells.

Flow cytometry

Each cell was washed in phosphate-buffered saline twice. After the cell count was adjusted to 1.5 × 10^6, cells were analyzed using an FACSCalibur cytometer (Becton Dickinson, Heidelberg, Germany). To confirm the double-transduction of MLL-AF4 and each Flt3 cDNA, we used FACS Vantage SE (Becton Dickinson). Individual transduced cells were sorted with FACS Vantage SE and used in this experiment.

RT-PCR analysis of MLL-AF4 and Flt3 cDNA from expression vectors

We confirmed the messenger RNA (mRNA) expression of each gene using RT-PCR. We also verified Flt3-TKD D835Y mutation by *EcoRV* restriction enzyme treatment. We confirmed the presence or absence of mutation by *EcoRV* restriction enzyme treatment after PCR amplification. For PCR amplification, Takara Ex-Taq (Takara, Shiga, Japan) was used. These primer sequences and PCR conditions are available upon request.

Expression analysis by Western blotting

To confirm protein expression and Flt3 phosphorylation, Western blotting was performed, as described previously [20]. The isolate was transferred to Immobilon P membranes (Millipore Bedford, MA, USA), and reacted with a primary antibody (MLLT2;

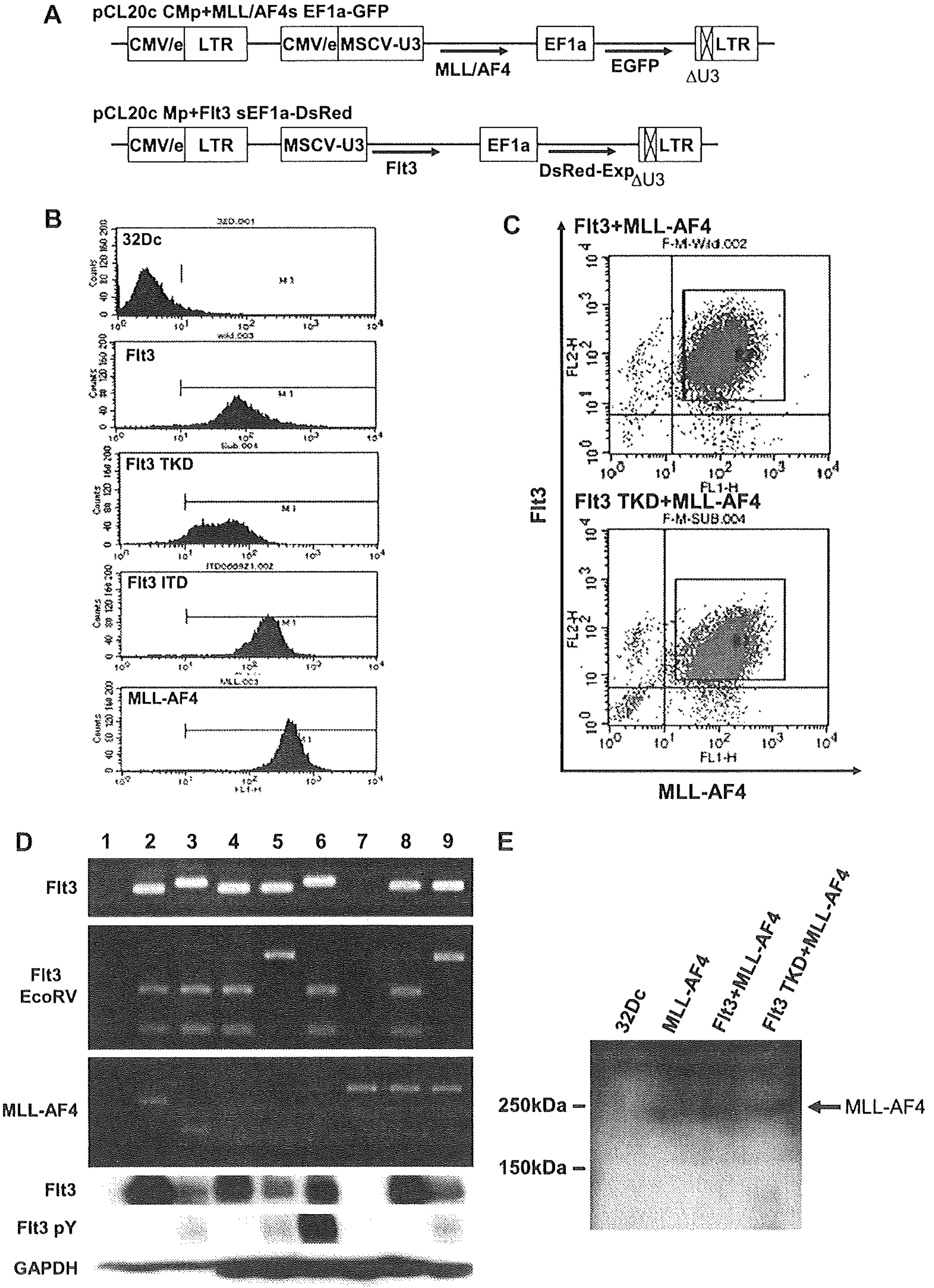

Figure 1. Preparation of 32Dc cells expressing Fms-like tyrosine kinase 3 (Flt3) mutations and/or MLL-AF4. (**A**) The lentiviral vector constructions are shown. CMV/e = cytomegalovirus immediate early enhancer; MSCV-U3 = murine stem cell virus (MSCV) LTR-U3 promoter; LTR = HIV-1 LTR; EF1a = human elongation factor 1 α promoter. (**B**) We analyzed the expressions of individual genes transduced to 32Dc cells by flow cytometry (FACSCalibur cytometer). MLL-AF4 was confirmed based on enhanced green fluorescent protein (EGFP) expression. Flt3 was confirmed based on DsRed expression. (**C**) We verified double transduction by flow cytometry (FACS Vantage SE). (**D**) 1: 32Dc, 2: RS411, 3: MV411, 4: Flt3, 5: Flt3-TKD, 6: Flt3-ITD, 7: MLL-AF4, 8: Flt3 + MLL-AF4, 9: Flt3-TKD + MLL-AF4. The three upper rows represent reverse transcription polymerase chain reaction (RT-PCR) analysis. Concerning Flt3, the PCR product of internal tandem duplication (ITD) increased by 30 bp (insertion volume) (lanes 3 and 6). In Flt3 tyrosine kinase domain (TDK) mutation, the PCR product is not cut due to the loss of the EcoRV enzyme cleavage site-specific to wild-type Flt3 (lanes 5 and 9). The break point site differed among RS411, MV411, and MLL-AF4, which we cloned. The break-point site of RS411 was present in MLL exon10 and AF4 exon4. Its PCR product measured 543 bp. The break-point site of MV411 was present in MLL exon9 and AF4 exon5. Its PCR product measured 366 bp. The break-point site of our cloning MLL-AF4 was present in MLL exon11 and AF4 exon4. Its PCR product measured 657 bp. The 3 lower rows represent Western blotting. Flt3 pY indicates the phosphorylation of Flt3 Tyr591. (**E**) This represents the expression of MLL-AF4 chimeric protein (240 kDa) on Western blotting. An arrow indicates MLL-AF4 chimeric protein.

1:1,000, calcyclin [H-55]; 1:500, GAPDH-loading control; 1:10,000, others; 1:2000).

Analysis of cell growth

The 32Dc cells expressing MLL-AF4 and/or Flt3 constructs were washed twice with phosphate-buffered saline and resuspended in RPMI-1640 with 10% fetal calf serum alone or supplemented with IL-3 (2 ng/mL). The viable cells, determined by trypan blue exclusion, were counted every day until day 10. Cells were seeded at a concentration of 1×10^5 cells/mL and cell numbers were adjusted every day.

Experiments involving apoptosis induction in the absence of IL-3 and the G-CSF–related induction of differentiation

After each cell was washed in phosphate-buffered saline twice, cell culture was performed in the absence of IL-3. 32Dc cells were stained with allophycocyanin, Annexin-V, and propidium iodide (Becton Dickinson) on day 3, and the other cells on day 7. We analyzed apoptosis induction using an FACSCalibur cytometer. Similarly, in the absence of IL-3, 10 ng/mL G-CSF was added, and cell culture was performed. We analyzed whether G-CSF induces cell differentiation into granulocytes until day 9. At 1×10^2 cells/μL, cytospin specimens were prepared, and Wright-Giemsa staining was performed before microscopy.

Clonal growth in methylcellulose

To analyze clonal growth, 1 mL MethoCult H4230 (StemCell Technologies Inc., Vancouver, BC, Canada) was plated on a 35-mm culture dish in the presence of IL-3 (2 ng/mL), or cytokine absence. Stably transfected 32Dc cells expressing MLL-AF4 and/or Flt3 constructs were seeded at a concentration of 1×10^4 cells/dish. The colonies were photographed and counted on day 7. Results shown are representative of one of at least three independent experiments per construct.

Gene expression profiling experiments by GeneChip expression and real-time PCR analyses

Mouse genome-wide gene expression was examined using the Mouse Genome 430 2.0 probe array (GeneChip; Affymetrix, Santa Clara, CA, USA), which contains the oligonucleotide probe set for approximately 34,000 full-length genes and expressed sequence tags, according to manufacturer's protocol (Affymetrix).

Portions of unamplified cDNA were subjected to PCR with SYBR Green PCR Core Reagents (PE Applied Biosystems, Foster City, CA, USA). Incorporation of the SYBR Green dye into the PCR products was monitored in real time with an ABI PRISM 7700 sequence detection system (PE Applied Biosystems), thereby allowing determination of the threshold cycle at which the exponential amplification of PCR products begins. The threshold cycle values for cDNAs corresponding to the GAPDH and target genes were used to calculate the abundance of the target transcripts relative to that of GAPDH mRNA. The oligonucleotide primers are available upon request.

siRNA transfection

Nineteen-nucleotide, single-stranded RNAs directed against the fusion sequence of S100A6 were chemically synthesized (Takara, Shiga, Japan). The sense and antisense sequences are available upon request. As negative control siRNAs, we used the mismatch control SNC1 (Takara, Shiga, Japan). Exponentially growing cells were concentrated to 10^6 cells/mL in culture medium, and 100 μL cell suspension was pipetted into a 4-mm electroporation cuvette. Immediately before the electroporation step, siRNAs were added, yielding a final concentration of 200 nM. Electroporation was performed with a Fischer electroporator (Fischer, Heidelberg, Germany) using a rectangle pulse of 330 V for 10 ms. After incubating for 15 minutes at room temperature, cells were diluted 20-fold with culture medium and incubated at 37°C and 5% CO_2.

Statistical analysis

The cell growth assay, colony-formation assay, differentiation assay, and gene expression assay were analyzed by Student's *t*-test, assuming unequal variances and two-tailed distributions. Data from every experiment represent the mean value ± standard deviation of triplicates.

Results

Expression and phosphorylation analysis of Flt3 and MLL-AF4 in 32Dc cells Flt3, Flt3-TKD, Flt3-ITD, and MLL-AF4 were individually transduced to a mouse IL-3–dependent cell line, 32Dc (Fig. 1B). In addition, MLL-AF4 was transduced to wild-type Flt3 or Flt3-TKD–transduced cells to prepare Flt3 + MLL-AF4 and Flt3-TKD + MLL-AF4, respectively (Fig. 1C). Concerning Flt3 protein expression, Flt3-TKD expression was lower than wild-type Flt3 and Flt3-ITD expressions (Fig. 1D, lanes 5 and 9). Due to its low protein expression, the phosphorylation activity of TKD was less marked than that of ITD (Fig. 1D, lanes 5 and 9). We compared the transduction efficiency of each lentivirus vector in 32Dc cells (Fig. 1B and 1C). Transduction efficiency of the TKD-expressing vector was slightly lower, suggesting that protein expression was low.

Flt3-ITD and MLL-AF4 + Flt3-TKD-expressing 32Dc cells resist induction of apoptosis after IL-3 deprivation

In the absence of IL-3, 32Dc cells in which each construction was introduced were cultured to examine cell proliferation. The Flt3-ITD–expressing 32Dc ($32Dc^{Flt3\text{-}ITD}$) and Flt3-TKD + MLL-AF4–expressing 32Dc ($32Dc^{Flt3\text{-}TKD+MLL\text{-}AF4}$) cells proliferated (Fig. 2). The Flt3-TKD–expressing 32Dc ($32Dc^{Flt3\text{-}TKD}$) cells also proliferated until day 4, as observed in the previously mentioned cells. However, finally, they did not proliferate independently (Fig. 2). The $32Dc^{Flt3\text{-}TKD+MLL\text{-}AF4}$ cells more slowly proliferated compared to the $32Dc^{Flt3\text{-}ITD}$ cells in the absence of IL-3. This was also similar in the presence of IL-3.

MLL-AF4 was frequent in pro – B ALL. To determine the leukemogenesis of MLL-AF4 in the lymphoid lineage, we also transduced Flt3-TKD and MLL-AF4 to a mouse IL-3–dependent cell line, Baf3 ($Baf3^{Flt3\text{-}TKD+MLL\text{-}AF4}$). However, $Baf3^{Flt3\text{-}TKD+MLL\text{-}AF4}$ did not proliferate independently in the absence of IL-3 (data not shown). MLL-AF4