Human Genome Center

Laboratory of Genome Database
ゲノムデータベース分野

Professor	Minoru Kanehisa, Ph.D.	教 授	理学博士	金 久 實
Assistant Professor	Toshiaki Katayama, M.Sc.	助 教	理学修士	片 山 俊 明
Assistant Professor	Shuichi Kawashima, M.Sc.	助 教	理学修士	川 島 秀 一

Since the completion of the Human Genome Project, high-throughput experimental projects have been initiated for uncovering genomic information in an extended sense, including transcriptomics, proteomics, metabolomics, glycomics, chemical genomics, and metagenomics. We are developing bioinformatics technologies to integrate and interpret such large-scale datasets, especially for medical and pharmaceutical applications.

1. KEGG DISEASE and KEGG DRUG

Minoru Kanehisa

KEGG is a database of biological systems that integrates genomic, chemical, and systemic functional information. It is widely used as a reference knowledge base for understanding higher-order functions and utilities of the cell, the organism, and the ecosystem from genomic information. This Laboratory is responsible for the applied areas of KEGG, especially in medical and pharmaceutical sciences. We consider diseases as perturbed states of the molecular system that operates the cell and the organism, and drugs as perturbants to the molecular system. We develop a disease information resource, KEGG DISEASE (http://www.genome.jp/kegg/disease/), which is intended for use by computational analysis rather than just for humans to read and understand. When the detail of the molecular system is relatively well characterized, we draw KEGG pathway maps. When the detail is not known but disease genes are identified, we create KEGG DISEASE entries, each of which contains a list of known disease genes and other relevant molecules including environmental factors, diagnostic markers, and therapeutic drugs. The list simply defines the membership to the underlying molecular system, but is useful for integrated analysis with large-scale datasets. We also develop a comprehensive drug information resource, KEGG DRUG (http://www.genome.jp/kegg/drug/), containing chemical structures and/or chemical components of all prescription and OTC drugs in Japan, and most prescription drugs in the USA and Europe. In KEGG DRUG we also capture knowledge on two types of molecular networks. One is the interaction network of drugs with target molecules, metabolizing enzymes, transporters, pharmacogenomic markers, other drugs, and the pathways involving all these molecules. The other is the chemical structure transformation network in the history of drug development where drug structures have been continuously modified by medicinal chemists.

2. KEGG OC: Automatic assignments of orthologs and paralogs in complete genomes

Toshiaki Katayama, Shuichi Kawashima, Akihiro Nakaya[1] and Minoru Kanehisa: [1]Graduate

School of Frontier Sciences, The University of Tokyo.

The increase in the number of complete genomes has provided clues to gain useful insights to understand the evolution of the gene universe. Among the KEGG suites of databases, the GENES database contains more than 5.6 million genes from over 1,300 organisms as of January 2011. Sequence similarities among these genes are calculated by all-against-all SSEARCH comparison and stored in the SSDB database. Based on those databases, the ORTHOLOGY database has been manually constructed to store the relationships among the genes sharing the same biological function. However, in this strategy, only the well known functions can be used for annotation of newly added genes, thus the number of annotated genes is limited. To overcome this situation, we have developed a fully automated procedure to find candidate orthologous clusters including those without any functional annotation. The method is based on a graph analysis of the SSDB database, treating genes as nodes and the Smith-Waterman sequence similarity scores as edge weights. The cluster is found by our heuristic method for finding quasi-cliques, but the SSDB graph is too large to perform quasi-clique finding at a time. Therefore, we introduce a hierarchy (evolutionary relationship) of organisms and treat the SSDB graph as a nested graph. The automatic decomposition of the SSDB graph into a set of quasi-cliques results in the KEGG OC (Ortholog Cluster) database. We have built a system that performs automatic update of KEGG OC, which can be run on a weekly basis. As a result, we obtained 959,333 clusters including 595,004 singleton clusters from 5,631,381 protein coding genes. Among them, 5,704 clusters were shared across kingdoms and other clusters were kingdom specific. The automatic classification of our ortholog clusters is largely consistent with the manually curated ORTHOLOGY database. A web interface to search and browse genes in clusters is made available at http://oc.hgc.jp/.

3. EGENES: A database for expressed sequence tag indices of plant species

Shuichi Kawashima, Yuki Moriya[1], Toshiaki Tokimatsu[1], Susumu Goto[1] and Minoru Kanehisa: [1]Institute for Chemical Science, Kyoto University.

EGENES is a knowledge-based database for efficient analysis of plant expressed sequence tags (ESTs). It links plant genomic information to higher order functional information in KEGG. The genomic information in EGENES is a collection of EST contigs constructed from assembled plant ESTs by using EGassembler. The EST indices are automatically annotated with the KEGG Orthology identifiers (K numbers) by KEGG Automatic Annotation Server (KAAS). Currently, EGENES contains 3,170,203 sequence catalogues in 80 plants, among which 21% have assigned K numbers. EGENES is available at http://www.genome.jp/kegg/catalog/org_list2.html

4. KEGG API: SOAP/WSDL interface for the KEGG system

Shuichi Kawashima, Toshiaki Katayama and Minoru Kanehisa

KEGG is a suite of databases and associated software, integrating our current knowledge of molecular interaction/reaction pathways and other systemic functions (PATHWAY and BRITE databases), information about the genomic space (GENES database), and information about the chemical space (LIGAND databases). To facilitate large-scale applications of the KEGG system programmatically, we have been developing and maintaining the KEGG API as a stable SOAP/WSDL based web service. The KEGG API is available at http://www.genome.jp/kegg/soap/.

5. BioRuby: Bioinformatics software for the Ruby programming language

Naohisa Goto[1], Pjotr Prins[2], Mitsuteru Nakao[3], Raoul Bonnal[4], Jan Aerts[5], and Toshiaki Katayama: [1]Research Institute for Microbial Diseases, Osaka University, [2]Database Center for Life Science, ROIS, [3]Groningen Bioinformatics Centre, University of Groningen, The Netherlands, [4]Integrative Biology Program, Fondazione Istituto Nazionale di Genetica Molecolare, Italy, [5]Katholieke Universiteit Leuven, Belgium.

The BioRuby software toolkit contains a comprehensive set of free development tools and libraries for bioinformatics and molecular biology, written in the Ruby programming language. BioRuby has components for sequence analysis, pathway analysis, protein modelling and phylogenetic analysis; it supports many widely used data formats and provides easy access to databases, external programs and public web services, including BLAST, KEGG, GenBank, MEDLINE and GO. BioRuby comes with a tutorial, documentation and an interactive environment, which can be used in the shell, and in the web browser. BioRuby is free and open source soft-

ware, made available under the Ruby license. BioRuby runs on all platforms that support Ruby, including Linux, Mac OS X and Windows. And, with JRuby, BioRuby runs on the Java Virtual Machine. The source code is available from http://www.bioruby.org/.

6. TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services

Toshiaki Katayama, Mitsuteru Nakao[1] and Toshihisa Takagi[2,3]: [1]Database Center for Life Science, ROIS, [2]National Bioscience Database Center, JST, [3]Graduate School of Frontier Sciences, The University of Tokyo.

Web services have become widely used in bioinformatics analysis, but there exist incompatibilities in interfaces and data types, which prevent users from making full use of a combination of these services. Therefore, we have developed the TogoWS service to provide an integrated interface with advanced features. In the TogoWS REST (REpresentative State Transfer) API (application programming interface), we introduce a unified access method for major database resources through intuitive URIs that can be used to search, retrieve, parse and convert the database entries. The TogoWS SOAP API resolves compatibility issues found on the server and client-side SOAP implementations. The TogoWS service is freely available at: http://togows.dbcls.jp/.

7. TogoDB: Instantly publish your research material as a public database

Toshiaki Katayama, Mitsuteru Nakao[1] and Toshihisa Takagi[2,3]: [1]Database Center for Life Science, ROIS, [2]National Bioscience Database Center, JST, [3]Graduate School of Frontier Sciences, The University of Tokyo.

Supplemental materials are often provided as separate files downloadable from the publisher's site along with the publication of a journal article. However, these data are not fully utilized since they are not available in the form of regular biological databases and hard to find by the popular Web search engines. TogoDB is a simple and intuitive database system to publish tabular formatted data instantly on the Web. Users can upload their research materials to the TogoDB through the simple web interface and the data will be made available as a fully functional database in a minute or two. TogoWS is an integrated and uniformed interface for the major bioinformatics web services and also provides REST API for the contents in the TogoDB. Recently, we extended TogoDB and TogoWS to be used as a consolidated platform for the Semantic Web by adding a metadata editor and an automatic Resource Description Framework (RDF) dumper. This system fills the gap between user's data and major public databases to deliver effective variations in the Linked Data.

8. Characterization of parasite specific genes by ortholog gene clusters and the Semantic Web technologies

Toshiaki Katayama, Shuichi Kawashima, Junichi Watanabe, Yutaka Suzuki[1], Sumio Sugano[1] and Minoru Kanehisa: [1]Graduate School of Frontier Sciences, The University of Tokyo.

Characterization of a set of genes is one of the most demanded tasks in bioinformatics. Usually, extensive sequence similarity search is performed and the functions of query sequences are inferred from the annotation of highly similar sequences. For more accurate inference, ortholog clusters and protein motif information are often added for further investigation. Ortholog cluster is useful to infer the phylogenetic distribution of the protein family and motifs provides molecular evidence of the functional domains in combination. In this way, information integration of related data sources is gaining in importance to make the annotation more reliable or to discover hidden relations among data sets. However the incompatible identifiers and data types have prevented from integrating heterogeneous data seamlessly among life-science databases. Recently, the Semantic Web technologies are being accepted to resolve this situation. We applied these methods to characterize a set of genes from non-lab strains of *Plasmodium falciparum* and other parasite species. Clinical samples are taken at the hospital in Manado, Indonesia and RNA sequencing was performed to find over-expressed malarial genes. As a result, we obtained ~5% of total genes which are highly expressed. Among them, we found 20 *P. falciparum* specific, 48 *Plasmodium* specific and 43 Apicomplexa specific genes. Many of them turned out to possess parasite specific domains such as surface antigen and exported protein with unknown function. We also discovered that some of them are changed 7-folds in their gene expression levels among samples which can be candidates for the drug target.

9. Analysis of EST sequences from the house dust mite, *Dermatophagoides farinae*

Shuichi Kawashima, Junichi Watanabe and Minoru Kanehisa

The house dust mite feed on skin scales (dander and scurf) and other organic detritus, such as bacteria, spores and feathers and produce 20 to 30 small fecal particles, which are highly allergic. Thus it is important that we expand our knowledge of the mites to develop effective allergy panels or vaccines. Angus *et al*. have initiated a project to sequence ESTs of *Dermatophagoides farinae* and other mite species associated with allergic diseases. Wakaguri *et al*. have also sequenced ESTs of *D. farinae* which was first established in 1968 and has been maintained using the method developed by Sasa *et al*. The produced sequences are available at the FullMite database. In this study we analyzed the relationships between known mite allergens and the EST sequences derived from the FullMite database.

After removing low quality regions and clipping vector sequences from the initial 23,040 sequences, a total of 21,005 sequences is retained. Then we assembled pair-end sequences if they were aligned with a strict condition (more than 500 bit score and 99% sequence identity in the aligned region). Clustering and assembling with the CAP3 software resulted in 1,717 contigs and 3,368 singletons. The distribution of the number of ESTs included in each contig showed that known allergens were widely distributed through the variety of gene expression levels. As expected, the major mite allergens such as Der f 1 or Der f 2 were highly expressed in our data. On the other hand there were also known allergens which were lower expressed. It is known at least 20% of adults allergic to mite have poor reactivity to the group 1 and group 2 allergens. Furthermore new mite allergens have still been characterized today (e.g. Der f 22). Thus there could be unknown mite allergens in our data and we will try to screen potential allergens which are still uncharacterized.

10. Analysis of a tardigrade proteome with the Gene Ontology

Shuichi Kawashima, Toshiaki Katayama and Minoru Kanehisa

Kunieda *et al*. have been sequencing and analyzing the genome of extremotolerant tardigrade, Rammazzottius cf. varieornatus, YOKOZUNA-1. Currently, they produced about 15,000 hypothetical protein coding genes by using *ab initio* method implemented in SNAP software. Tardigrades form the phylum Tardigrada which is considered as a member of the superphylum Ecdysozoa. However the definitive phylogenetic position is still unclear. To understand the ecdysozoans specific molecular functions, we compared the Gene Ontology terms annotated to the proteome with those of *Drosphila melanogaster*, *Daphnia pulex* and *Caenorhabditis elegans* as representative Ecdysozoas and found the 2,241 GO temrs common to the four ecdysozoans. Interestingly, only six ontologies were found in the proteomes of four species exclusively. Those are rhabdomere (GO:0016028), cation channel complex (GO:0034703), septate junction (GO:0005918), signal recognition particle (GO:0048500), fusom (GO:0045169) and spectrosome (GO:0045170).

11. Functional genome annotation of a tardigrade by KEGG MODULE database

Toshiaki Katayama, Shuichi Kawashima and Minoru Kanehisa

Functional annotation of genes is usually performed based on sequence similarities to genes of other organisms. However, it is difficult to find out characteristics of the genome only based on the each annotated gene. Therefore, overall signature of a genome is often described with the help of the Gene Ontology (GO) or KEGG PATHWAY databases. Assigning GO terms to genes, we can grasp an outline of functional categories covered by a set of genes encoded in a genome. Reconstructing KEGG PATHWAY will reveals metabolic and regulatory pathways that the target organism may utilize. The problem here is the granularity. Functional categories of the GO and covering area of each KEGG PATHWAY map are rather coarse-grained to find key differences among organisms. Meanwhile, KEGG MODULE is a database of functional units in the pathways which are highly conserved among organisms. We are running a genome project of a tardigrade, which is expected to explain its phylogenetic position and to make functional annotation of the genome. Therefore, we introduced the KEGG MODULE to see conserved and non-conserved modules in the tardigrade genome compared with closely-related species. As a result, we found 184 conserved modules in tardigrade which are slightly larger than that of *C. elegans* (162), *D. pulex* (181) and *D. melanogaster* (164). Among them, we found several module candidates describing evolution of the metabolic pathways in animal. For example, M00029 (urea

cycle) which is missing in nematodes, partially conserved (cytosolic portion) in arthropods, and fully conserved in vertebrates (cytosolic and mitochondrial portions). Tardigrade had only one enzyme in this module (in cytosolic portion) showing the same pattern with some protists. As the number of sequenced genomes increases very rapidly, KEGG MODULE combined with a phylogenetic profile will be a powerful tool to elucidate characteristics of those genomes efficiently.

12. HiGet and SSS: Search engines for the large-scale biological databases

Toshiaki Katayama, Shuichi Kawashima, Kazuhiro Ohi[1], Kenta Nakai[2] Minoru Kanehisa: [1]Hitatchi Ltd., [2]Laboratory of Functional Analysis In Silico, Institute of Medical Science, The University of Tokyo.

The number of entries in biological databases is exponentially increasing year by year. For example, there were 10,106,023 entries in the GenBank database in the year 2000, which has now grown to 141,537,514 (Release 181+daily updates). In order for such a vast amount of data to be searched at a high speed, we have developed a high performance database entry retrieval system, named HiGet. For this purpose, the system is constructed on the HiRDB, a commercial ORDBMS (Object-oriented Relational Database Management System) developed by Hitachi, Ltd. HiGet can perform full text search on various biological databases including GenBank, RefSeq, UniProt, Prosite, OMIM and PDB. Additional advantage of the HiGet system is the capability of a field specific search, which enables users to narrow down the number of results, especially useful for collecting sequences of their specific needs. We have also developed a sequence similarity search (SSS) service to find homologous sequences with various algorithms including BLAST, FASTA, SSEARCH, TRANS, and EXONERATE. This variety of options is unique among the public services and users can select an appropriate method to search similar sequences according to their query. Because algorithms such as TRANS and EXONERATE are highly time consuming, the SSS service is backended by the distributed computing environment with the Sun Grid Engine in our super computer system. HiGet and SSS services are available at http://higet.hgc.jp/ and http://sss.hgc.jp/ respectively.

Publications

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research* 38: D355-D360, 2010

Katayama, T., Nakao, M., Takagi, T. TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services. *Nucleic Acids Research*, 38: W706-W711, 2010

Katayama, T., Arakawa, K., Nakao, M., Ono, K., Aoki-Kinoshita, K.F., Yamamoto, Y., Yamaguchi, A., Kawashima, S., Chun, H.W., Aerts, J., Aranda, B., Barboza, L.H., Bonnal, R.J., Bruskiewich, R., Bryne, J.C., Fernández, J.M., Funahashi, A., Gordon, P.M., Goto, N., Groscurth, A., Gutteridge, A., Holland, R., Kano, Y., Kawas, E.A., Kerhornou, A., Kibukawa, E., Kinjo, A.R., Kuhn, M., Lapp, H., Lehvaslaiho, H., Nakamura, H., Nakamura, Y., Nishizawa, T., Nobata, C., Noguchi, T., Oinn, T.M., Okamoto, S., Owen, S., Pafilis, E., Pocock, M., Prins, P., Ranzinger, R., Reisinger, F., Salwinski, L., Schreiber, M., Senger, M., Shigemoto, Y., Standley, D.M., Sugawara, H., Tashiro, T., Trelles, O., Vos, R.A., Wilkinson, M.D., York, W., Zmasek, C.M., Asai, K., Takagi, T. The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows. *Journal of Biomedical Semantics*, 1: 8, 2010

Goto, N., Prins, P., Nakao, M., Bonnal, R., Aerts, J., Katayama, T. BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics*, 26: 2617-2619, 2010

Yamanishi, Y., Kotera, M., Kanehisa, M., Goto, S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 26: i246-i256, 2010

Erguner, B., Hattori, M., Goto, S., Kanehisa, M. Characterizing common substructures of ligands for GPCR protein subfamilies. *Genome Informatics* 24: 31-41, 2010

Kotera, M., Kobayashi, T., Hattori, M., Tokimatsu, T., Goto, S., Mihara, H., Kanehisa, M. Comprehensive genomic analysis of sulfur-relay pathway genes. *Genome Informatics* 24: 104-115, 2010

Mizutani, S., Tanaka, M., Wheelock, C., Kanehisa, M., Goto, S. Phylogenetic analysis of lipid mediator GPCRs. *Genome Informatics* 24: 116-126, 2010

Nishimura, Y., Tokimatsu, T., Kotera, M., Goto, S., Kanehisa, M. Genome-wide analysis of

plant UGT family based on sequence and substrate information. *Genome Informatics* 24: 127-138, 2010

Human Genome Center

Laboratory of DNA Information Analysis
Laboratory of Sequence Data Analysis
DNA情報解析分野
シークエンスデータ情報処理分野

Professor	Satoru Miyano, Ph.D.	教　授	理学博士	宮　野　　悟
Associate Professor	Seiya Imoto, Ph.D.	准教授	博士(数理学)	井　元　清　哉
Assistant Professor	Masao Nagasaki, Ph.D.	助　教	博士(理学)	長　崎　正　朗
Project Assistant Professor	Yoshinori Tamada, Ph.D.	特任助教	博士(情報学)	玉　田　嘉　紀
Project Assistant Professor	Teppei Shimamura, Ph.D.	特任助教	博士(工学)	島　村　徹　平
Associate Professor	Tetsuo Shibuya, Ph.D.	准教授	博士(理学)	渋　谷　哲　朗
Lecturer	Rui Yamaguchi, Ph.D.	講　師	博士(理学)	山　口　　類

The recent advances in biomedical research have been producing large-scale, ultra-high dimensional, ultra-heterogeneous data. Due to these post-genomic research progresses, our current mission is to create computational strategy for systems biology and medicine towards translational bioinformatics. With this mission, we have been developing computational methods for understanding life as system and applying them to practical issues in medicine and biology.

1. Gene Network Analysis

a. Inferring dynamic gene networks under varying conditions for transcriptomic network comparison

Teppei Shimamura, Seiya Imoto, Rui Yamaguchi, Masao Nagasaki, Satoru Miyano

Elucidating the differences between cellular responses to various biological conditions or external stimuli is an important challenge in systems biology. Many approaches have been developed for reverse-engineering a cellular system, called gene network, from time series microarray gene expression data in order to understand a transcriptomic response under a condition of interest. Comparative topological analysis has also been applied based on the gene networks inferred independently from each of the multiple time series datasets under varying conditions to find critical differences between these networks. However, these comparisons often lead to misleading results, because each network contains considerable noise due to the limited length of the time series. With this motivation, we developed an integrated approach for inferring multiple gene networks from time series expression data under varying conditions. To the best of our knowledge, our approach is the first reverse-engineering method that is intended for transcriptomic network comparison between varying conditions. Furthermore, we developed a state-of-the-art parameter estimation method, relevance-weighted recursive elastic net, for providing higher precision and recall than existing reverse-engineering methods. We analyze experimental data of MCF-7 human breast cancer

cells stimulated by epidermal growth factor or heregulin with several doses and provide novel biological hypotheses through network comparison. The software NETCOMP is available at http : / / bonsai . ims . u-tokyo . ac . jp /~shima / NETCOMP/.

b. Collocation-based sparse estimation for constructing dynamic gene networks

Teppei Shimamura, Seiya Imoto, Masao Nagasaki, Mai Yamauchi[13], Rui Yamaguchi, André Fujita, Yoshinori Tamada, Noriko Gotoh[13], Satoru Miyano

One of the open problems in systems biology is to infer dynamic gene networks describing the underlying biological process with mathematical, statistical and computational methods. The first-order difference equation-based models such as dynamic Bayesian networks and vector autoregressive models were used to infer time-lagged relationships between genes from time-series microarray data. However, two primary problems greatly reduce the effectiveness of current approaches. The first problem is the tacit assumption that time lag is stationary. The second is the inseparability between measurement noise and process noise (unmeasured disturbances that pass through time process). To address these problems, we developed a stochastic differential equation model for inferring continuous-time dynamic gene networks under the situation in which both of the process noise and the observation noise exist. We devised a collocation-based sparse estimation for simultaneous parameter estimation and model selection in the model. The collocation-based approach requires considerably less computational effort than traditional methods in ordinary stochastic differential equation models. We also incorporated various biological knowledge easily to refine the estimation accuracy with this method. The results using simulated data and real time-series expression data of human primary small airway epithelial cells demonstrate that this approach outperformed competing approaches and could provide significant genes influenced by Gefitinib.

c. Gene set-based module discovery decodes *cis*-regulatory codes governing diverse gene expression across human multiple tissues

Atsushi Niida, Seiya Imoto, Rui Yamaguchi, Masao Nagasaki, Saatoru Miyano

Decoding transcriptional programs governing transcriptomic diversity across human multiple tissues is a major challenge in bioinformatics. To address this problem, a number of computational methods have focused on *cis*-regulatory codes driving overexpression or underexpression in a single tissue as compared to others. On the other hand, we recently proposed a different approach to mine *cis*-regulatory codes: starting from gene sets sharing common *cis*-regulatory motifs, the method screens for expression modules based on expression coherence. However, both approaches seem to be insufficient to capture transcriptional programs that control gene expression in a subset of all samples. Especially, this limitation would be serious when analyzing multiple tissue data. To overcome this limitation, we developed a new module discovery method termed BEEM (Biclusering-based Extraction of Expression Modules) in order to discover expression modules that are functional in a subset of tissues. We showed that, when applied to expression profiles of human multiple tissues, BEEM finds expression modules missed by two existing approaches that are based on the coherent expression and the single tissue-specific differential expression. From the BEEM results, we obtained new insights into transcriptional programs controlling transcriptomic diversity across various types of tissues. This study introduces BEEM as a powerful tool for decoding regulatory programs from a compendium of gene expression profiles.

d. Model-free unsupervised gene set screening based on information enrichment in expression profiles

Atsushi Niida, Seiya Imoto, Rui Yamaguchi, Masao Nagasaki, André Fujita, Teppei Shimamura, Satoru Miyano

A number of unsupervised gene set screening methods have recently been developed for search of putative functional gene sets based on their expression profiles. Most of the methods statistically evaluate whether the expression profiles of each gene set are fit to assumed models, e.g. co-expression across all samples or a subgroup of samples. However, it is possible that they fail to capture informative gene sets whose expression profiles are not fit to the assumed models. To overcome this limitation, we developed a model-free unsupervised gene set screening method, Matrix Information Enrichment Analysis (MIEA). Without assuming any specific models, MIEA screens gene sets based on information richness of their expression profiles. We extensively compared the performance of MIEA to those of other unsupervised gene set screen-

ing methods, using various types of simulated and real data. The benchmark tests demonstrated that MIEA can detect singular expression profiles that the other methods fail to find, and performs broadly well for various types of input data. Taken together, this study introduces MIEA as a broadly applicable gene set screening tool for mining regulatory programs from transcriptome data.

e. Gene regulatory network clustering for graph layout based on microarray gene expression data

Kaname Kojima, Seiya Imoto, Masao Nagasaki, Satoru Miyano

We developed a statistical model realizing simultaneous estimation of gene regulatory network and gene module identification from time series gene expression data from microarray experiments. Under the assumption that genes in the same module are densely connected, this method detects gene modules based on the variational Bayesian technique. The model can also incorporate existing biological prior knowledge such as protein subcellular localization. We applied our model to the time series data from a synthetically generated network and verified the effectiveness of the proposed model. This model is also applied to the time series microarray data from HeLa cell. Detected gene module information gave the great help on drawing the estimated gene network.

f. Optimal search on clustered structural constraint for learning Bayesian network structure

Kaname Kojima, Eric Perrier, Seiya Imoto, Satoru Miyano

We studied the problem of learning an optimal Bayesian network in a constrained search space; skeletons are compelled to be subgraphs of a given undirected graph called the super-structure. The previously derived constrained optimal search (COS) remains limited even for sparse super-structures. To extend its feasibility, we developed a method to divide the super-structure into several clusters and perform an optimal search on each of them. Further, to ensure acyclicity, we introduced the concept of ancestral constraints (ACs) and derive an optimal algorithm satisfying a given set of ACs. Finally, we theoretically derived the necessary and sufficient sets of ACs to be considered for finding an optimal constrained graph. Empirical evaluations demonstrated that our algorithm can learn optimal Bayesian networks for some graphs containing several hundreds of vertices, and even for super-structures having a high average degree (up to four), which is a drastic improvement in feasibility over the previous optimal algorithm. Learnt networks were shown to largely outperform state-of-the-art heuristic algorithms both in terms of score and structural hamming distance.

g. A fast and robust statistical test based on likelihood ratio with Bartlett correction to identify Granger causality between gene sets

André Fujita, Kaname Kojima, Alexandre G. Patriota[1], João Ricardo Sato[2], Patricia Severino[3], Satoru Miyano: [1]University of São Paulo, [2]Universidade Federal do ABC, [3]Albert Einstein Research and Education Institute

We developed a likelihood ratio test (LRT) with Bartlett correction in order to identify Granger causality between sets of time series gene expression data. The performance of the proposed test is compared to a previously published bootstrap-based approach. LRT is shown to be significantly faster and statistically powerful even within non-Normal distributions. An R package named gGranger containing an implementation for both Granger causality identification tests is also provided: http://dnagarden.ims.u-tokyo.ac.jp/afujita/en/doku.php?id=ggranger

h. Comparison of gene expression profiles produced by CAGE, Illumina microarray and Real Time RT-PCR

André Fujita, Masao Nagasaki, Seiya Imoto, Ayumu Saito, Emi Ikeda, Teppei Shimamura, Rui Yamaguchi, Yoshihide Hayashizaki[4], Satoru Miyano: [4]RIKEN Yokohama Institute

Several technologies are currently used for gene expression profiling, such as Real Time RT-PCR, microarray and CAGE (Cap Analysis of Gene Expression). CAGE is a recently developed method for constructing transcriptome maps and it has been successfully applied to analyzing gene expressions in diverse biological studies. The principle of CAGE has been developed to address specific issues such as determination of transcriptional starting sites, the study of promoter regions and identification of new transcripts. We made both quantitative and qualitative comparisons among three major gene expression quantification techniques, namely: CAGE, illumina microarray and Real Time RT-

PCR, by showing that the quantitative values of each method are not interchangeable, however, each of them has unique characteristics which render all of them essential and complementary. Understanding the advantages and disadvantages of each technology will be useful in selecting the most appropriate technique for a determined purpose.

i. Identification of Granger causality between gene sets

André Fujita, João Ricardo Sato[2], Kaname Kojima, Luciana Rodrigues Gomes[1], L.R., Mari Cleide Sogayar[1], Satoru Miyano

Wiener and Granger have introduced an intuitive concept of causality (Granger causality) between two variables which is based on the idea that an effect never occurs before its cause. Later, Geweke generalized this concept to a multivariate Granger causality, i.e. *n* variables Granger-cause another variable. Although Granger causality is not "effective causality" in the Aristothelic sense, this concept is useful to infer directionality and information flow in observational data. Granger causality is usually identified by using VAR (Vector Autoregressive) models due to their simplicity. In the last few years, several VAR-based models were presented in order to model gene regulatory networks. We generalized the multivariate Granger causality concept in order to identify Granger causalities between sets of gene expressions, i.e. whether a set of *n* genes Granger-causes another set of *m* genes, aiming at identifying the flow of information between gene networks (or pathways). The concept of Granger causality for sets of variables is presented. Moreover, a method for its identification with a bootstrap test is proposed. This method is applied in simulated and also in actual biological gene expression data in order to model regulatory networks. This concept may be useful for the understanding of the complete information flow from one network or pathway to the other, mainly in regulatory networks. Linking this concept to graph theory, sink and source can be generalized to node sets. Moreover, hub and centrality for sets of genes can be defined based on total information flow. Another application is in annotation, when the functionality of a set of genes is unknown, but this set is Granger-caused by another set of genes which is well studied. Therefore, this information may be useful to infer or construct some hypothesis about the unknown set of genes.

j. Granger causality in systems biology: modeling gene networks in time series microarray data using vector autoregressive models

André Fujita, Patricia Severino[3], João Ricardo Sato[2], Miyano, S.

Understanding the molecular biological processes underlying disease onset requires a detailed description of which genes are expressed at which time points and how their products interact in so-called cellular networks. High-throughput technologies, such as gene expression analysis using DNA microarrays, have been extensively used with this purpose. As a consequence, mathematical methods aiming to infer the structure of gene networks have been proposed in the last few years. Granger causality-based models are among them, presenting well established mathematical interpretations to directionality at the edges of the regulatory network. Here, we describe the concept of Granger causality and explore recent advances and applications in gene expression regulatory networks by using extensions of Vector Autoregressive models.

k. Discovering functional gene pathways associated with cancer heterogeneity via sparse supervised learning

Shuichi Kawano, Teppei Shimamura, Atsuhi Niida, Seiya Imoto, Rui Yamaguchi, Masao Nagasaki, Ryo Yoshida[5], Cristin Print[6], Satoru Miyano: [5]Institute of Statistical Mathematics [6]University of Auckland

We developed a statistical method for uncovering gene pathways that characterize cancer heterogeneity. To incorporate knowledge of the pathways into the model, we define a set of activities of pathways from microarray gene expression data based on the sparse probabilistic principal component analysis. A pathway activity logistic regression model is then formulated for cancer phenotype. To select pathway activities related to binary cancer phenotypes, we use the elastic net for the parameter estimation and derive a model selection criterion for selecting tuning parameters included in the model estimation. Our method can also reverse-engineer gene networks based on the identified multiple pathways that enables us to discover novel gene-gene associations relating with the cancer phenotypes. We illustrated the whole process of the proposed method through the analysis of breast cancer gene expression data.

l. Identifying hidden confounders in gene networks by Bayesian networks

Tomoya Higashigaki[7], Kaname Kojima, Rui Yamaguchi, Masato Inoue[7], Seiya Imoto, Satoru Miyano: [7]Waseda University

For estimating gene networks from microarray gene expression data, we developed a statistical method for quantification of the hidden confounders in gene networks, which were possibly removed from the set of genes on the gene networks or are novel biological elements that are not measured by microarrays. Due to high computational cost of the structural learning of Bayesian networks and the limited source of the microarray data, it is usual to perform gene selection prior to the estimation of gene networks. Therefore, there exist missing genes that decrease accuracy and interpretability of the estimated gene networks. The proposed method can identify hidden confounders based on the conflicts of the estimated local Bayesian network structures and estimate their ideal profiles based on the proposed Bayesian networks with hidden variables with an EM algorithm. From the estimated ideal profiles, we can identify genes which are missing in the network or suggest the existence of the novel biological elements if the ideal profiles are not significantly correlated with any expression profiles of genes. To the best of our knowledge, this research is the first study to theoretically characterize missing genes in gene networks and practically utilize this information to refine network estimation.

2. Pathway Modeling, Simulation and Analysis

a. Cell Illustrator 4.0: A computational platform for systems biology

Masao Nagasaki, Ayumu Saito, Euna Jeong, Chen Li, Kaname Kojima, Emi Ikeda, Satoru Miyano

Cell Illustrator is a software platform for Systems Biology that uses the concept of Petri net for modeling and simulating biopathways. It is intended for biological scientists working at bench. The latest version of Cell Illustrator 4.0 uses Java Web Start technology and is enhanced with new capabilities, including: automatic graph grid layout algorithms using ontology information; tools using Cell System Markup Language (CSML) 3.0 and Cell System Ontology 3.0; parameter search module; high-performance simulation module; CSML database management system; conversion from CSML model to programming languages (FORTRAN, C, C++, Java, Python and Perl); import from SBML, CellML, and BioPAX; and, export to SVG and HTML. Cell Illustrator employs an extension of hybrid Petri net in an object-oriented style so that biopathway models can include objects such as DNA sequence, molecular density, 3D localization information, transcription with frame-shift, translation with codon table, as well as biochemical reactions.

b. Time-dependent structural transformation analysis to high-level Petri net model with active state transition diagram

Chen Li, Masao Nagasaki, Ayumu Saito, Satoru Miyano

Investigating the dynamic features of current computational models promises a deeper understanding of complex cellular processes. This leads us to develop a method that utilizes structural properties of the model over all simulation time steps. Further, user-friendly overviews of dynamic behaviors can be considered to provide a great help in understanding the variations of system mechanisms. We developed a novel method for constructing and analyzing a so-called active state transition diagram (ASTD) by using time-course simulation data of a high-level Petri net. Our method includes two new algorithms. The first algorithm extracts a series of subnets (called temporal subnets) reflecting biological components contributing to the dynamics, while retaining positive mathematical qualities. The second one creates an ASTD composed of unique temporal subnets. ASTD provides users with concise information allowing them to grasp and trace how a key regulatory subnet and/or a network changes with time. The applicability of our method is demonstrated by the analysis of the underlying model for circadian rhythms in *Drosophila.* Building ASTD is a useful means to convert a hybrid model dealing with discrete, continuous and more complicated events to finite time-dependent states. Based on ASTD, various analytical approaches can be applied to obtain new insights into not only systematic mechanisms but also dynamics.

c. On determining delay time of transitions for Petri net based signaling pathways by introducing stochastic decision rules

Yoshimasa Miwa[8], Chen Li, Qi-Wei Ge[8], Hiroshi Matsuno[8], Satoru Miyano: [8]Yamaguchi University

Parameter determination is important in mod-

eling and simulating biological pathways including signaling pathways. Parameters are determined according to biological facts obtained from biological experiments and scientific publications. However, such reliable data describing detailed reactions are not reported in most cases. This prompted us to develop a general methodology of determining the parameters of a model in the case of that no information of the underlying biological facts is provided. In this study, we used the Petri net approach for modeling signaling pathways, and developed a method to determine firing delay times of transitions for Petri net models of signaling pathways by introducing stochastic decision rules. Petri net technology provides a powerful approach to modeling and simulating various concurrent systems, and recently has been widely accepted as a description method for biological pathways. Our method enables to determine the range of firing delay time which realizes smooth token flows in the Petri net model of a signaling pathway. The availability of this method has been confirmed by the results of an application to the interleukin-1 induced signaling pathway.

d. An efficient biological pathway layout algorithm combining grid-layout and spring embedder for complicated cellular location information

Kaname Kojima, Masao Nagasaki, Satoru Miyano

We developed a new grid-layout algorithm based on the spring embedder algorithm that can handle location information and provide layouts with harmonized appearance. In grid-layout algorithms, the mapping of nodes to grid points that minimizes a cost function is searched. By imposing positional constraints on grid points, location information including complex shapes can be easily considered. Our layout algorithm includes the spring embedder cost as a component of the cost function. We further extended the layout algorithm to enable dynamic update of the positions and sizes of compartments at each step. The new spring embedder-based grid-layout algorithm and a spring embedder algorithm were applied to three biological pathways; endothelial cell model, Fas-induced apoptosis model, and *C. elegans* cell fate simulation model. From the positional constraints, all the results of our algorithm satisfy location information, and hence, more comprehensible layouts were obtained as compared to the spring embedder algorithm. From the comparison of the number of crossings, the results of the grid-layout-based algorithm tend to contain more crossings than those of the spring embedder algorithm due to the positional constraints.

3. Data Assimilation for Systems Biology

e. DA 1.0: parameter estimation of biological pathways using data assimilation approach

Chuan Hock Koh[9], Masao Nagasaki, Ayumu Saito, Limsoon Wong[10], Satoru Miyano: [9]National University of Singapore and Human Genome Center, Institute of Medical Science, University of Tokyo, [10]National University of Singapore

Data assimilation (DA) is a computational approach that estimates unknown parameters in a pathway model using time-course information. Particle filtering, the underlying method used, is a well-established statistical method that approximates the joint posterior distributions of parameters by using sequentially generated Monte Carlo samples. We released the Java-based software (DA 1.0) with an intuitive and user-friendly interface to allow users to carry out parameters estimation using DA. DA 1.0 was developed using Java and thus would be executable on any platform installed with JDK 6.0 (not JRE 6.0) or later. DA 1.0 is freely available for academic users and can be launched or downloaded from http://da.csml.org.

f. Phosphoproteomics-based modeling defines the regulatory mechanism underlying aberrant EGFR signaling

Shinya Tasaki[11], Masao Nagasaki, M., Hiroko Kozuka-Hata[11], Kentaro Semba[12], Noriko Gotoh[13], Seisuke Hattori[14], Jun-ichiro Inoue[15], Tadashi, Yamamoto[15], Satoru Miyano, Sumio Sugano[16], Masaaki Oyama: [11]Medical Proteomics Laboratory, Institute of Medical Science, University of Tokyo, [12]Department of Life Science and Medical Bio-Science, Waseda University, [13]Division of Systems Biomedical Technology, Institute of Medical Science, University of Tokyo, [14]Department of Biochemistry, School of Pharmaceutical Sciences, Kitasato University, [15]Department of Cancer Biology, Institute of Medical Science, University of Tokyo, [16]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo

Mutation of the epidermal growth factor receptor (EGFR) results in a discordant cell signaling, leading to the development of various dis-

eases. However, the mechanism underlying the alteration of downstream signaling due to such mutation has not yet been completely understood at the system level. Here, we report a phosphoproteomics-based methodology for characterizing the regulatory mechanism underlying aberrant EGFR signaling using computational network modeling. Our phosphoproteomic analysis of the mutation at tyrosine 992 (Y992), one of the multifunctional docking sites of EGFR, revealed network-wide effects of the mutation on EGF signaling in a time-resolved manner. Computational modeling based on the temporal activation profiles enabled us to not only rediscover already-known protein interactions with Y992 and internalization property of mutated EGFR but also further gain model-driven insights into the effect of cellular content and the regulation of EGFR degradation. Our kinetic model also suggested critical reactions facilitating the reconstruction of the diverse effects of the mutation on phosphoproteome dynamics. Our integrative approach provided a mechanistic description of the disorders of mutated EGFR signaling networks, which could facilitate the development of a systematic strategy toward controlling disease-related cell signaling.

4. Next-Generation Sequencer Data Analysis

a. International network of cancer genome projects

International Cancer Genome Consortium

The International Cancer Genome Consortium (ICGC) was launched to coordinate large-scale cancer genome studies in tumours from 50 different cancer types and/or subtypes that are of clinical and societal importance across the globe. Systematic studies of more than 25,000 cancer genomes at the genomic, epigenomic and transcriptomic levels will reveal the repertoire of oncogenic mutations, uncover traces of the mutagenic influences, define clinically relevant subtypes for prognosis and therapeutic management, and enable the development of new cancer therapies. In this project, we developed a next-generation sequence data analysis pipeline for the supercomputer system of Human Genome Center.

b. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing

Akihiro Fujimoto[17], Hidewaki Nakagawa[17], Naoya Hosono[17], Kaoru Nakano[17], Tetsuo Abe[17], Keith A. Boroevich[17], Masao Nagasaki, Rui Yamaguchi, Tetsuo Shibuya, Michiaki Kubo[17], Satoru Miyano, Yusuke Nakamura, Tatsuhiko Tsunoda[17]: [17]Center for Genomic Medicine, RIKEN

We report the analysis of a Japanese male using high-throughput sequencing to×40 coverage. More than 99% of the sequence reads were mapped to the reference human genome. Using a Bayesian decision method, we identified 3,132,608 single nucleotide variations (SNVs). Comparison with six previously reported genomes revealed an excess of singleton nonsense and nonsynonymous SNVs, as well as singleton SNVs in conserved non-coding regions. We also identified 5,319 deletions smaller than 10 kb with high accuracy, in addition to copy number variations and rearrangements. De novo assembly of the unmapped sequence reads generated around 3 Mb of novel sequence, which showed high similarity to non-reference human genomes and the human herpesvirus 4 genome. Our analysis suggests that considerable variation remains undiscovered in the human genome and that whole-genome sequencing is an invaluable tool for obtaining a complete understanding of human genetic variation. In this research, we developed a next-generation sequence data analysis pipeline for the supercomputer system of Human Genome Center.

5. Algorithms for Protein Structures

a. Geometric suffix tree: Indexing protein 3-D structures

Tetsuo Shibuya

Protein structure analysis is one of the most important research issues in the post-genomic era, and faster and more accurate index data structures for such 3-D structures are highly desired for research on proteins. This article proposes a new data structure for indexing protein 3-D structures. For strings, there are many efficient indexing structures such as suffix trees, but it has been considered very difficult to design such sophisticated data structures against 3-D structures like proteins. Our index structure is based on the suffix tree and is called the geometric suffix tree. By using the geometric suffix tree for a set of protein structures, we can exactly search for all of their substructures whose RMSDs (root mean square deviations) or URMSDs (unit-vector root mean square deviations) to a given query 3-D structure are not larger than a given bound. Though there are $O(N^2)$ substructures in a structure of size N, our

data structure requires only $O(N)$ space for indexing all the substructures. We propose an $O(N^2)$ construction algorithm for it, while a naive algorithm would require $O(N^3)$ time to construct it. Moreover we propose an efficient search algorithm. Experiments show that we can search for similar structures much faster than previous algorithms if the RMSD threshold is not larger than 1 Å. The experiments also show that the construction time of the geometric suffix tree is practically almost linear to the size of the database, when applied to a protein structure database.

b. Searching protein 3-D structures in faster than linear time

Tetsuo Shibuya

Searching for similar structures from a three-dimensional (3-D) structure database of proteins is one of the most important problems in post-genomic computational biology. To compare two structures, we ordinarily use a measure called the root mean square deviation (RMSD) as the similarity measure. We consider a very fundamental problem of finding all the substructures whose RMSDs to the query are within some given threshold, from a 3-D structure database. The problem also appears in many other fields, such as computer vision and robotics. In this article, we propose the first algorithm that runs in faster than linear time on average. Our new algorithm runs in average-case $O(m+N/m(1-\varepsilon))$, where N is the database size, m is the query length, and epsilon is an arbitrary small constant such that $0<\varepsilon<1$. It is a significant improvement over previous algorithms on the problem, considering that the best known worst-case time complexity of the problem is $O(N \log m)$, and the best known average-case (expected) time complexity of the problem was $O(N)$.

6. Pandemic Control Simulation

a. When should we intervene to control the 2009 influenza A(H1N1) pandemic?

Hiroki Sato[18], Haruka Nakada[19], Rui Yamaguchi, Seiya Imoto, Satoru Miyano, Masahiro Kami[19]: [18]Department of Medical Informatics, National Defense Medical College Hospital, [19]Division of Social Communication System for Advanced Clinical Research, Institute of Medical Science, University of Tokyo

We simulated the early phase of the 2009 influenza A(H1N1) pandemic and assessed the effectiveness of public health interventions in Japan. We show that the detection rate of border quarantine was low and the timing of the intervention was the most important factor involved in the control of the pandemic, with the maximum reduction in daily cases obtained after interventions started on day 6 or 11. Early interventions were not always effective.

Publications

1. Do, J.H., Nagasaki, M., Miyano, S. The systems approach to the prespore-specific activation of sigma factor SigF in *Bacillus subtilis*. Biosystems. 100: 178-184, 2010.
2. Ferreira, C.E., Miyano, S., Stadler, P.F. (Eds.) Advances in Bioinformatics and Computational Biology, Lecture Notes in Computer Science Vol. 6268, Springer, 2010.
3. Fujimoto, A., Nakagawa, H., Hosono, N., Nakano, K., Abe, G., Boroevich, K.A., Nagasaki, M., Yamaguchi, R., Shibuya, T., Kubo, M., Miyano, S., Nakamura, Y., Tsunoda, T. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nature Genetics. 42: 931-936, 2010.
4. Fujita, A., Kojima, K., Patriota, A.G., Sato, J. R., Severino, P., Miyano, S. A fast and robust statistical test based on likelihood ratio with Bartlett correction to identify Granger causality between gene sets. Bioinformatics. 26(18): 2349-2351, 2010.
5. Fujita, A., Nagasaki, M., Imoto, S., Saito, A., Ikeda, E., Shimamura, T., Yamaguchi, R., Hayashizaki, Y., Miyano, S. Comparison of gene expression profiles produced by CAGE, Illumina microarray and Real Time RT-PCR. Genome Informatics. 24: 56-68, 2010.
6. Fujita, A., Sato, J.R., Demasi, M.A.A., Miyano, S., Sogayar, M.C., Ferreira, C.E. An introduction to time-varying connectivity estimation for gene regulatory networks. "Medical Biostatistics for Complex Diseases" (Frank Emmert-Streib; Matthias Dehmer. (Eds.)). Weinheim, Germany, Wiley VCH Verlag, 205-230, 2010.
7. Fujita, A., Sato, J.R., Kojima, K., Gomes, L.R., Sogayar, M.C., Miyano, S. Identification of Granger causality between gene sets. J. Bioinformatics and Computational Biology. 8(4): 679-701, 2010.
8. Fujita, A., Severino, P., Sato, J.R., Miyano, S. Granger causality in systems biology: mod-

eling gene networks in time series microarray data using vector autoregressive models. Lecture Notes in Bioinformatics. 6268: 13-24, 2010.
9. Higashigaki,T., Kojima, K., Yamaguchi, R., Inoue, M., Imoto, S., Miyano, S. Identifying hidden confounders in gene networks by Bayesian networks. Proc. 10th IEEE Bioinformatics and Bioengineering. 168-173, 2010.
10. International Cancer Genome Consortium, Hudson, T.J. et al. International network of cancer genome projects. Nature. 464(7291): 993-998, 2010.
11. Kaufmann, K., Nagasaki, M., Jáuregui, R. Modelling the molecular interactions in the flower developmental network of Arabidopsis thaliana. In Silico Biol. 10: 0008, 2010.
12. Kawano, S., Shimamura, T., Niida, A., Imoto, S., Yamaguchi, R., Nagasaki, M., Yoshida, R., Print, C., Miyano, S. Discovering functional gene pathways associated with cancer heterogeneity via sparse supervised learning. Proc. IEEE 10th International Symposium on Bioinformatics & Bioengineering. 253-258, 2010.
13. Koh, C.H., Nagasaki, M., Saito, A., Wong, L., Miyano, S. DA 1.0: parameter estimation of biological pathways using data assimilation approach. Bioinformatics. 26(14): 1794-1796, 2010.
14. Kojima, K., Imoto, S., Nagasaki, M., Miyano, S. Gene regulatory network clustering for graph layout based on microarray gene expression data. Genome Informatics. 24: 84-95, 2010.
15. Kojima, K., Nagasaki, M., Miyano, S. An efficient biological pathway layout algorithm combining grid-layout and spring embedder for complicated cellular location information. BMC Bioinformatics. 11: 335, 2010.
16. Kojima, K., Perrier, E., Imoto, S., Miyano, S. Optimal search on clustered structural constraint for learning Bayesian network structure. J. Machine Learning Research.11: 285-310, 2010.
17. Li, C., Nagasaki, M., Saito, A., Miyano, S. Time-dependent structural transformation analysis to high-level Petri net model with active state transition diagram. BMC Systems Biology. 4: 39, 2010.
18. Mitou, N., Matsuno, H., Miyano, S., Inouye, S. Essential role of Ror gene in the interaction of feedback loops in mammalian circadian clocks. "Modeling in Systems Biology-The Petri Net Approach" (Koch, I., Reisig, W., Schreiber, F. (Eds.)), Springer, 281-306, 2010.
19. Miwa, Y., Li, C., Ge, Q.W., Matsuno, H., Miyano, S. On determining delay time of transitions for Petri net based signaling pathways by introducing stochastic decision rules, In Silico Biol. 10: 0004, 2010.
20. Nagasaki, M., Saito, A., Jeong, E., Li, C., Kojima, K., Ikeda, E., Miyano, S., Cell Illustrator 4.0: A computational platform for systems biology. In Silico Biol. 10: 0002, 2010.
21. Niida, A., Imoto, S., Yamaguchi, R., Nagasaki, M., Miyano, S. Gene set-based module discovery decodes *cis*-regulatory codes governing diverse gene expression across human multiple tissues. PLoS ONE. 5(6): e 10910, 2010.
22. Niida, A., Imoto, S., Yamaguchi, R., Nagasaki, M., Fujita, A., Shimamura, T., Miyano, S. Model-free unsupervised gene set screening based on information enrichment in expression profiles. Bioinformatics. 26(24): 3090-3097, 2010.
23. Saito, A., Nagasaki, M., Miyano, S. Hybrid functional Petri net with extension for dynamic pathway modeling. "Modeling in Systems Biology-The Petri Net Approach" (Koch, I., Reisig, W., Schreiber, F. (Eds.)), Springer, 101-120, 2010.
24. Sato, H., Nakada, H., Yamaguchi, R., Imoto, S., Miyano, S., Kami, M. When should we intervene to control the 2009 influenza A(H1 N1) pandemic? Euro Surveill. 7: 15(1), pii: 19455, 2010.
25. Shibuya, T. Searching protein 3-D structures in faster than linear time. J. Comput. Biol. 17 (4): 593-602, 2010.
26. Shibuya, T. Searching protein 3-D structures in linear time. J. Comput. Biol. 17(3): 203-219, 2010.
27. Shibuya, T. Geometric suffix tree: Indexing protein 3-D structures. J. ACM. 57(3): 1-17, 2010.
28. Shibuya, T. Fast hinge detection algorithms for flexible protein structures. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 7(2): 333-341, 2010.
29. Shibuya, T., Jansson, J., Sadakane. K., Linear-time protein 3-D structure searching with insertions and deletions. BMC Algorithms for Molecular Biology. 5: 7, 2010.
30. Shimamura, T., Imoto, S., Yamaguchi, R., Nagasaki, M., Miyano, S. Inferring dynamic gene networks under varying conditions for transcriptomic network comparison. Bioinformatics. 26(8): 1064-1072, 2010.
31. Shimamura, T., Imoto, S., Nagasaki, M., Yamauchi, M., Yamaguchi, R., Fujita, A., Tamada, Y., Gotoh, N., Miyano, S. Collocation-based sparse estimation for constructing dynamic gene networks. Genome Informatics. 24: 164-178, 2010.
32. Sogawa, Y., Shimizu, S, Hyvarinen, A.,

Washio, T., Shimamura, T., Imoto, S. Discovery of exogenous variables in data with more variables than observations. Proc. 20th International Conference on Artificial Neural Networks, 67-76, 2010.
33. Tasaki, S., Nagasaki, M., Kozuka-Hata, H., Semba, K., Gotoh, N., Hattori, S., Inoue, J., Yamamoto, T., Miyano, S., Sugano, S., Oyama, M. Phosphoproteomics-based modeling defines the regulatory mechanism underlying aberrant EGFR signaling. PLoS ONE. 5(11), e13926, 2010.
34. Yamaguchi, R., Imoto, S., Miyano, S. Network-based predictions and simulations by biological state space models: Search for drug mode of action. J. Computer Science and Technology. 25(1): 13-153, 2010.
35. Yuji, K., Matsumura, T., Miyano, S., Tsuchiya, R., Kami, M. Human papillomavirus vaccine coverage. Lancet. 376(9738): 329-330, 2010.

Human Genome Center

Laboratory of Molecular Medicine
Laboratory of Genome Technology
ゲノムシークエンス解析分野
シークエンス技術開発分野

Professor	Yusuke Nakamura, M.D., Ph.D.	教　授	医学博士	中　村　祐　輔
Associate Professor	Koichi Matsuda, M.D., Ph.D.	准教授	医学博士	松　田　浩　一
Assistant Professor	Ryuji Hamamoto, Ph.D.	助　教	理学博士	浜　本　隆　二
Assistant Professor	Hitoshi Zembutsu, M.D., Ph.D.	助　教	医学博士	前　佛　　均

The major goal of our group is to identify genes of medical importance, and to develop new diagnostic and therapeutic tools. We have been attempting to isolate genes involving in carcinogenesis and also those causing or predisposing to various diseases as well as those related to drug efficacies and adverse reactions. By means of technologies developed through the genome project including a high-resolution SNP map, a large-scale DNA sequencing, and the cDNA microarray method, we have isolated a number of biologically and/or medically important genes, and are developing novel diagnostic and therapeutic tools.

1. Genes playing significant roles in human cancer

Koichi Matsuda, Yataro Daigo, Hidewaki Nakagawa, Ryuji Hamamoto, Hitoshi Zembutsu, Chikako Fukukawa, Jae-Hyun Park, Yosuke Harada, Masahiko Ajiro, Jung-Won Kim, Koji Ueda, Nguyen Minh-Hue, Junkichi Koinuma, Daiki Miki, Ken Masuda, Masato Aragaki, Takashi Fujitomo, Hideto Oshita, Satoko Uno, Yoichiro Kato, Su-Youn Chung, Lianhua Piao, Chizu Tanikawa, Cui Ri, Hamdi Mbarek, Vinod Kumar, Osman W Mohammed, Yuji Urabe, Jiaying Lin, Zhenzhong Deng, Martha Espinosa, Motoko Unoki, Masanori Yoshimatsu, Shinya Hayami, Hyun-Soo Cho, Goji Toyokawa, Tadashi Takawa, Reem Abdelrahim Ibrahim, Seham Elgazzar, Mitsuko Nakashima, Kang Daechun, Cha Pei Chieng, Low Siew Kee, and Yusuke Nakamura

(1) Lung cancer

Dickkopf-1

Dickkopf-1 (DKK1) is an inhibitor of Wnt/beta-catenin signaling that is overexpressed in most lung and esophageal cancers. Here, we show its utility as a serum biomarker for a wide range of human cancers, and we offer evidence favoring the potential application of anti-DKK1 antibodies for cancer treatment. Using an original ELISA system, high levels of DKK1 protein were found in serologic samples from 906 patients with cancers of the pancreas, stomach, liver, bile duct, breast, and cervix, which also showed elevated expression levels of DKK1. Additionally, anti-DKK1 antibody inhibited the invasive activity and the growth of cancer cells in vitro and suppressed the growth of engrafted tumors in vivo. Tumor tissues treated with anti-DKK1 displayed significant fibrotic changes and

a decrease in viable cancer cells without apparent toxicity in mice. Our findings suggest DKK1 as a serum biomarker for screening against a variety of cancers, and anti-DKK1 antibodies as potential theranostic tools for diagnosis and treatment of cancer.

WDHD1 (WD repeat and high-mobility group box DNA binding protein 1)

To identify novel biomarkers and therapeutic targets for lung and esophageal cancers, we screened for genes that were overexpressed in a large proportion of lung and esophageal carcinomas using a cDNA microarray representing 27,648 genes or expressed sequence tags. A gene encoding *WDHD1*, a WD repeat and high-mobility group box DNA binding protein 1, was selected as a candidate. Tumor tissue microarray analyses covering 267 archival non-small cell lung cancers and 283 esophageal squamous cell carcinomas (ESCC) revealed that positive WDHD1 immunostaining was associated with a poor prognosis for patients with non-small cell lung cancer ($P = 0.0403$) as well as ESCC ($P = 0.0426$). Multivariate analysis indicated it to be an independent prognostic factor for ESCC ($P = 0.0104$). Suppression of WDHD1 expression with small interfering RNAs effectively suppressed lung and esophageal cancer cell growth. In addition, induction of the exogenous expression of WDHD1 promoted the growth of mammalian cells. AKT1 kinase seemed to phosphorylate and stabilize the WDHD1 protein in cancer cells. WDHD1 expression is likely to play an important role in lung and esophageal carcinogenesis as a cell cycle regulator and a downstream molecule in the phosphoinositide 3-kinase/AKT pathway, and that WDHD1 is a candidate biomarker and a promising therapeutic target for cancer.

CDCA5 (cell division cycle associated 5)

We analyzed the gene expression profiles of clinical lung carcinomas using a cDNA microarray containing 27,648 genes or expressed sequence tags, and identified CDCA5 (cell division cycle associated 5) to be upregulated in the majority of lung cancers. Tumor tissue microarray analysis of 262 non-small cell lung cancer patients revealed that CDCA5 positivity was an independent prognostic factor for lung cancer patients. Suppression of CDCA5 expression with siRNAs inhibited the growth of lung cancer cells; concordantly, induction of exogenous expression of CDCA5 conferred growth-promoting activity in mammalian cells. We also found that extracellular signal-regulated kinase (ERK) kinase phosphorylated CDCA5 at Ser79 and Ser 209 in vivo. Exogenous expression of phosphomimicking CDCA5 protein whose Ser209 residue was replaced with glutamine acid further enhanced the growth of cancer cells. In addition, functional inhibition of the interaction between CDCA5 and ERK kinase by a cell-permeable peptide corresponding to a 20-amino-acid sequence part of CDCA5, which included the Ser 209 phosphorylation site by ERK, significantly reduced phosphorylation of CDCA5 and resulted in growth suppression of lung cancer cells. Our data suggest that transactivation of CDCA5 and its phosphorylation at Ser209 by ERK play an important role in lung cancer proliferation, and that the selective suppression of the ERK-CDCA5 pathway could be a promising strategy for cancer therapy.

(2) Pancreatic cancer

Involvement of TTLL4 Polyglutamylase in PELP1 Polyglutamylation and Chromatin Remodeling in Pancreatic Cancer Cells

Polyglutamylation is a new class of post-translational modification in which glutamate side chains are formed on proteins although its biological significance is not well known. Through our genome-wide gene-expression profile analysis of pancreatic ductal adenocarcinoma (PDAC) cells, we identified overexpression of TTLL4 (tubulin tyrosine ligase-like family member 4) in PDAC cells. Subsequent RT-PCR and northern-blot analyses confirmed its up-regulation in several PDACs. TTLL4 belongs to the TTLL family that was reported to have polyglutamylase activity. Knockdown of TTLL4 by shRNA in PDAC cells attenuated the growth of PDAC cells and exogenous introduction of TTLL4 enhanced the cell growth. We also found that TTLL4 expression was correlated with polyglutamylation levels of a glutamate-stretch region of PELP1 (proline, glutamate and leucine rich protein 1) that was shown to interact with various proteins such as histone H3, and be involved in several signaling pathways through its function as a scaffold protein. PELP1 polyglutamylation could influence to its interaction with histone H3 and affect histone H3 acetylation. We also identified the interaction of PELP1 with LAS1L and SENP3, components of the MLL1-WDR5 super-complex involving chromatin-remodelling. Our findings imply that TTLL4 could play important roles in pancreatic carcinogenesis through its polyglutamylase activity and subsequent coordination of chromatin remodeling, and might be a good molecular candidate for development of new therapeutic

strategies for pancreatic cancer.

C12orf48, termed PARP-1 binding protein (PARPBP), Enhances Poly(ADP-ribose) Polymerase-1 (PARP-1) Activity and Protects Pancreatic Cancer Cells from DNA Damage

To identify novel therapeutic targets for aggressive and therapy-resistant pancreatic cancer, we had previously performed expression profile analysis of pancreatic cancers using microarrays and found dozens of genes trans-activated in pancreatic ductal adenocarcinoma (PDAC) cells. Among them, this study focused on the characterization of a novel gene *C12orf48* whose overexpression in PDAC cells was validated by northern blot and immunohistochemical analyses. Its overexpression was observed in other aggressive and therapy-resistant malignancies as well. Knockdown of C12orf48 by siRNA in PDAC cells significantly suppressed their growth. Importantly, we demonstrated that C12 orf48 protein could directly interact with Poly (ADP-ribose) Polymerase-1 (PARP-1), one of the essential proteins in the repair of DNA damage, and positively regulate the poly(ADP-ribosyl) ation activity of PARP-1. Depletion of C12orf48 sensitized PDAC cells to agents causing DNA damage and also enhanced DNA damage-induced G2/M arrest through reduction of PARP-1 enzymatic activities. Hence, our findings implicate C12orf48, termed PARP-1 binding protein (PARPBP), or its interaction with PARP-1 to be a potential molecular target for development of selective therapy for pancreatic cancer.

(3) Prostate cancer

Association of a Novel Long Non-coding RNA in *8q24* with Prostate Cancer Susceptibility

Recent genome-wide association studies reported strong and reproducible associations of multiple genetic variants in a large "gene-desert" region of chromosome *8q24* with susceptibility to prostate cancer (PC). However, the causative or functional variants of these *8q24* loci and their biological mechanisms associated with PC susceptibility remain unclear and should to be investigated. Here, focusing on its most centromeric region (so-called Region 2: Chr 8: 128.14-128.28Mb) among the multiple PC loci on *8q24*, we performed fine mapping and resequencing of this critical region and identified SNPs between rs1456315 and rs7463708 (chr8: 128,173,119-128,173,237bp) to be most significantly associated with PC susceptibility ($P = 2.00 \times 10^{-24}$, OR=1.74, 95% CI=1.56-1.93). Importantly, we here show that this region was transcribed as a ~13-kb intron-less long non-coding RNA (ncRNA), termed as *PRNCR1* (*prostate cancer non-coding RNA 1*), and *PRNCR1* expression was up-regulated in some of PC cells as well as precursor lesion PINs. Knockdown of *PRNCR1* by siRNA attenuated the viability of PC cells and the transactivation activity of androgen receptor, which indicates that *PRNCR1* could be involved in prostate carcinogenesis possibly through androgen receptor activity. These findings could provide a new insight to understand the pathogenesis of genetic factors for PC susceptibility and prostate carcinogenesis.

(4) p53 target genes

XEDAR (X-linked ectodermal dysplasia receptor)

We recently identified X-linked ectodermal dysplasia receptor (XEDAR, also known as TNFRSF27 or EDA2R) as a direct p53 target that was frequently downregulated in colorectal cancer tissues due to its epigenetic alterations or through the p53 gene mutations. However, the role of the posttranslational regulation of XEDAR protein in colorectal carcinogenesis was not well clarified thus far. Here, we report that the extracellular NH(2) terminus of XEDAR protein was cleaved by a metalloproteinase and released into culture media. The remaining COOH-terminal membrane-anchored fragment was rapidly degraded through the ubiquitin-proteasome pathway. Interestingly, ectopic p53 expression also transactivated an XEDAR ligand, EDA-A2, together with XEDAR. Moreover, EDA-A2 blocked the cleavage of XEDAR and subsequently inhibited cell growth. We also found a missense mutation of the XEDAR gene in NCI-H716 colorectal cancer cells, which caused the translocation of XEDAR protein from cell membrane to cytoplasm. This mutation attenuated the growth-suppressive effect of XEDAR, indicating that membrane localization is critical for physiologic XEDAR function. Thus, our findings clearly revealed the crucial role of EDA-A2/XEDAR interaction in the p53-signaling pathway.

2. Pharmacogenetics

Lessons for pharmacogenomics studies: association study between CYP2D6 genotype and tamoxifen response.

We earlier reported a significant association between the cytochrome P450 2D6 (CYP2D6) genotype and the clinical outcome in 282 Japanese breast cancer patients receiving tamoxifen

monotherapy. Although many research groups have provided evidence indicating the CYP2D6 genotype as one of the strongest predictors of tamoxifen response, the results still remain controversial. We hypothesized that concomitant treatment was one of the causes of these controversial results. We then studied 167 breast cancer patients who received tamoxifen-combined therapy to evaluate the effects of concomitant treatment on the association analysis and observed no significant association between CYP2D6 genotype and recurrence-free survival (P=0.44, hazard ratio: 0.64, 95% confidential interval: 0.20-1.99 in patients with two variant alleles vs. patients without a variant allele). When we carried out two subgroup analyses for nodal status and tumor size, we observed a positive association between the CYP2D6 genotype and the clinical outcome only in patients who received tamoxifen monotherapy. This study explained a part of the discrepancies among the reported results.

3. Genome-wide association study

(1) cancer susceptible gene

Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations.

Lung cancer is the most common cause of death from cancer worldwide, and its incidence is increasing in East Asian and Western countries. To identify genetic factors that modify the risk of lung adenocarcinoma, we conducted a genome-wide association study in a Japanese cohort, with replication in two independent studies in Japanese and Korean individuals, in a total of 2,098 lung adenocarcinoma cases and 11,048 controls. The combined analyses identified two susceptibility loci for lung adenocarcinoma: TERT (rs2736100, combined $P=2.91\times10^{-11}$), odds ratio (OR)=1.27) and TP63 (rs 10937405, combined $P=7.26\times10^{-12}$), OR=1.31). Fine mapping of the region containing TP63 showed that a SNP (rs4488809) in intron 1 of TP63 showed the most significant association. Our results suggest that genetic variation in TP63 may influence susceptibility to lung adenocarcinoma in East Asian populations.

Genome-wide association study of pancreatic cancer in Japanese population.

Pancreatic cancer shows very poor prognosis and is the fifth leading cause of cancer death in Japan. Previous studies indicated some genetic factors contributing to the development and progression of pancreatic cancer; however, there are limited reports for common genetic variants to be associated with this disease, especially in the Asian population. We have conducted a genome-wide association study (GWAS) using 991 invasive pancreatic ductal adenocarcinoma cases and 5,209 controls, and identified three loci showing significant association (P-value$<5\times10^{-7}$) with susceptibility to pancreatic cancer. The SNPs that showed significant association carried estimated odds ratios of 1.29, 1.32, and 3.73 with 95% confidence intervals of 1.17-1.43, 1.19-1.47, and 2.24-6.21; P-value of 3.30×10^{-7}, $3.30\times10(-7)$, and 4.41×10^{-7}; located on chromosomes 6p25.3, 12p11.21 and 7q36.2, respectively. These associated SNPs are located within linkage disequilibrium blocks containing genes that have been implicated some roles in the oncogenesis of pancreatic cancer.

Common variant in 6q26-q27 is associated with distal colon cancer in Asian population

Colorectal cancer (CRC) is a multifactorial disease with both environmental and genetic factors contributing to its development. The incidence of CRC is increasing year by year in Japan. Patients with CRC in advanced stages have a poor prognosis, but detection of CRC at earlier stages can improve clinical outcome. Therefore, identification of epidemiologic factors that influence development of CRC would facilitate the prevention or early detection of disease.

To identify loci associated with CRC risk, we performed a genome-wide association study (GWAS) for CRC and sub-analyses by tumor location using 1,583 Japanese CRC cases and 1,898 controls. Subsequently, we conducted replication analyses using a total of 4,809 CRC cases and 2,973 controls including 225 Korean subjects with distal colon cancer and 377 controls.

We identified a novel locus on 6q26-q27 region (rs7758229 in SLC22A3, P=7.92×10-9, Odds ratio of 1.28) that was significantly associated with distal colon cancer. We also replicated the association between CRC and SNPs on 8q24 (rs6983267 and rs7837328, P=1.51×10-8 and 7.44×10-8, Odds ratios of 1.18 and 1.17, respectively). Moreover, we found cumulative effects of three genetic (rs7758229, rs6983267, and rs 4939827 in SMAD7) and one environmental factors (alcohol drinking) which appear to increase CRC risk approximately twofold.

We found a novel susceptible locus in SLC22 A3 that contributes to the risk of distal colon cancer in Asian population. These findings would further extend our understanding of the role of common genetic variants in CRC etiology.

(2) other diseases

A genome-wide association study identifies four susceptibility loci for keloid in the Japanese population.

Keloid is a dermal fibroproliferative growth that results from dysfunction of the wound healing processes. Through a multistage genome-wide association study using 824 individuals with keloid (cases) and 3,205 unaffected controls in the Japanese population, we identified significant associations of keloid with four SNP loci in three chromosomal regions: 1q41, 3q22.3-23 and 15q21.3. The most significant association with keloid was observed at rs873549 (combined $P=5.89\times10^{-23}$, odds ratio (OR)=1.77) on chromosome 1. Associations on chromosome 3 were observed at two separate linkage disequilibrium (LD) blocks: rs1511412 in the LD block including FOXL2 with $P=2.31\times10^{-13}$ (OR=1.87) and rs 940187 in another LD block with $P=1.80\times10^{-13}$ (OR=1.98). Association of rs8032158 located in NEDD4 on chromosome 15 yielded $P=5.96\times10^{-13}$ (OR=1.51). Our findings provide new insights into the pathophysiology of keloid formation.

A genome-wide association study identifies genetic variants in the CDKN2BAS locus associated with endometriosis in Japanese.

Although the pathogenesis of endometriosis is not well understood, genetic factors have been considered to have critical roles in its etiology. Through a genome-wide association study and a replication study using a total of 1,907 Japanese individuals with endometriosis (cases) and 5,292 controls, we identified a significant association of endometriosis with rs10965235 ($P=5.57\times10^{-12}$, odds ratio=1.44), which is located in CDKN2BAS on chromosome 9p21, encoding the cyclin-dependent kinase inhibitor 2B antisense RNA. By fine mapping, the SNP showing the strongest association was located in intron 16 of CDKN2BAS and was implicated in regulating the expression of p15, p16 and p14. A SNP, rs 16826658, in the LD block including WNT4 on chromosome 1p36, which is considered to play an important role in the development of the female genital tract, revealed a possible association with endometriosis ($P=1.66\times10^{-6}$, odds ratio=1.20). Our findings suggest that these regions are new susceptibility loci for endometriosis

Genome-wide association study of intracranial aneurysm identifies three new risk loci.

Saccular intracranial aneurysms are balloon-like dilations of the intracranial arterial wall; their hemorrhage commonly results in severe neurologic impairment and death. We report a second genome-wide association study with discovery and replication cohorts from Europe and Japan comprising 5,891 cases and 14,181 controls with approximately 832,000 genotyped and imputed SNPs across discovery cohorts. We identified three new loci showing strong evidence for association with intracranial aneurysms in the combined dataset, including intervals near RBBP 8 on 18q11.2 (odds ratio (OR)=1.22, $P=1.1\times10^{-12}$), STARD13-KL on 13q13.1 (OR=1.20, $P=2.5\times10^{-9}$) and a gene-rich region on 10q24.32 (OR=1.29, $P=1.2\times10^{-9}$). We also confirmed prior associations near SOX17 (8q11.23-q12.1; OR=1.28, $P=1.3\times10^{-12}$) and CDKN2A-CDKN2 B (9p21.3; OR=1.31, $P=1.5\times10^{-22}$). It is noteworthy that several putative risk genes play a role in cell-cycle progression, potentially affecting the proliferation and senescence of progenitor-cell populations that are responsible for vascular formation and repair

(3) Quantitative trait loci

Genome-wide association study of hematological and biochemical traits in a Japanese population.

We report genome-wide association studies for hematological and biochemical traits from approximately 14,700 Japanese individuals. We identified 60 associations for 8 hematological traits and 29 associations for 12 biochemical traits at genome-wide significance levels ($P<5\times10(-8)$). Of these, 46 associations were new to this study and 43 replicated previous reports. We compared these associated loci with those reported in similar GWAS in European populations. When the minor allele frequency was >10% in the Japanese population, 32 (94.1%) and 31 (91.2%) of the 34 hematological loci previously reported to be associated in a European population were replicated with P-values less than 0.05 and 0.01, respectively, and 31 (73.8%) and 27 (64.3%) of the 42 European biochemical loci were replicated.

Publications

1. Kiyotani, K., Mushiroda, T., Imamura, C.K., Hosono, N., Tsunoda, T., Kubo, M., Tanigawara, Y., Flockhart, D.A., Desta, Z., Skaar, T.C., Aki, F., Hirata, K., Takatsuka, Y., Okazaki, M., Ohsumi, S., Yamakawa, T., Sasa, M., Nakamura, Y., and Zembutsu, H. Significant effect of polymorphisms in CYP2 D6 and ABCC2 on clinical outcomes of adjuvant tamoxifen therapy for breast cancer patients. J Clin Oncol, *28*: 1287-1293, 2010.
2. Sato, N., Koinuma, J., Fujita, M., Hosokawa, M., Ito, T., Tsuchiya, E., Kondo, S., Nakamura, Y., and Daigo, Y. Activation of WD repeat and high-mobility group box DNA binding protein 1 in pulmonary and esophageal carcinogenesis. Clin Cancer Res, *16*: 226-239, 2010.
3. Wangsomboonsiri, W., Mahasirimongkol, S., Chantarangsu, S., Kiertiburanakul, S., Charoenyingwattana, A., Komindr, S., Thongnak, C., Mushiroda, T., Nakamura, Y., Chantratita, W., and Sungkanuparph, S. Association between HLA-B*4001 and lipodystrophy among HIV-infected patients from Thailand who received a stavudine-containing antiretroviral regimen. Clin Infect Dis, *50*: 597-604, 2010.
4. Miyazawa, M., Ohsawa, R., Tsunoda, T., Hirono, S., Kawai, M., Tani, M., Nakamura, Y., and Yamaue, H. Phase I clinical trial using peptide vaccine for human vascular endothelial growth factor receptor 2 in combination with gemcitabine for patients with advanced pancreatic cancer. Cancer Sci, *101*: 433-439, 2010.
5. Nuinoon, M., Makarasara, W., Mushiroda, T., Setianingsih, I., Wahidiyat, P.A., Sripichai, O., Kumasaka, N., Takahashi, A., Svasti, S., Munkongdee, T., Mahasirimongkol, S., Peerapittayamongkol, C., Viprakasit, V., Kamatani, N., Winichagoon, P., Kubo, M., Nakamura, Y., and Fucharoen, S. A genome-wide association identified the common genetic variants influence disease severity in beta0-thalassemia/hemoglobin E. Hum Genet, *127*: 303-314, 2010.
6. Nakahara, H., Sekiguchi, K., Hosono, N., Kubo, M., Takahashi, A., Nakamura, Y., and Kasai, K. Criterion values for multiplex SNP genotyping by the invader assay. Forensic Sci Int Genet, *4*: 130-136, 2010.
7. Nakahara, H., Hosono, N., Kitayama, T., Sekiguchi, K., Kubo, M., Takahashi, A., Nakamura, Y., Yamano, Y., and Kai, K. Automated SNPs typing system based on the Invader assay. Leg Med (Tokyo), *11 Suppl 1*: S111-114, 2009.
8. Sato, N., Koinuma, J., Ito, T., Tsuchiya, E., Kondo, S., Nakamura, Y., and Daigo, Y. Activation of an oncogenic TBC1D7 (TBC1 domain family, member 7) protein in pulmonary carcinogenesis. Genes Chromosomes Cancer, *49*: 353-367, 2010.
9. Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y., Nakamura, Y., and Kamatani, N. Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat Genet, *42*: 210-215, 2010.
10. Inoue, M., Senju, S., Hirata, S., Ikuta, Y., Hayashida, Y., Irie, A., Harao, M., Imai, K., Tomita, Y., Tsunoda, T., Furukawa, Y., Ito, T., Nakamura, Y., Baba, H., and Nishimura, Y. Identification of SPARC as a candidate target antigen for immunotherapy of various cancers. Int J Cancer, *127*: 1393-1403, 2010.
11. Maeda, S., Kobayashi, M.A., Araki, S., Babazono, T., Freedman, B.I., Bostrom, M.A., Cooke, J.N., Toyoda, M., Umezono, T., Tarnow, L., Hansen, T., Gaede, P., Jorsal, A., Ng, D.P., Ikeda, M., Yanagimoto, T., Tsunoda, T., Unoki, H., Kawai, K., Imanishi, M., Suzuki, D., Shin, H.D., Park, K.S., Kashiwagi, A., Iwamoto, Y., Kaku, K., Kawamori, R., Parving, H.H., Bowden, D.W., Pedersen, O., and Nakamura, Y.A single nucleotide polymorphism within the acetyl-coenzyme A carboxylase beta gene is associated with proteinuria in patients with type 2 diabetes. PLoS Genet, *6*: e1000842, 2010.
12. Park, J.H., Nishidate, T., Kijima, K., Ohashi, T., Takegawa, K., Fujikane, T., Hirata, K., Nakamura, Y., and Katagiri, T. Critical roles of mucin 1 glycosylation by transactivated polypeptide N-acetylgalactosaminyltransferase 6 in mammary carcinogenesis. Cancer Res, *70*: 2759-2769, 2010.
13. Mototani, H., Iida, A., Nakamura, Y., and Ikegawa, S. Identification of sequence polymorphisms in CALM2 and analysis of association with hip osteoarthritis in a Japanese population. J Bone Miner Metab, *28*: 547-553, 2010.
14. Prescott, N.J., Dominy, K.M., Kubo, M., Lewis, C.M., Fisher, S.A., Redon, R., Huang, N., Stranger, B.E., Blaszczyk, K., Hudspith, B., Parkes, G., Hosono, N., Yamazaki, K., Onnie, C.M., Forbes, A., Dermitzakis, E.T., Nakamura, Y., Mansfield, J.C., Sanderson, J., Hurles, M.E., Roberts, R.G., and Mathew, C. G. Independent and population-specific association of risk variants at the IRGM locus with Crohn's disease. Hum Mol Genet, *19*: 1828-1839, 2010.

15. Hayami, S., Yoshimatsu, M., Veerakumarasivam, A., Unoki, M., Iwai, Y., Tsunoda, T., Field, H.I., Kelly, J.D., Neal, D.E., Yamaue, H., Ponder, B.A., Nakamura, Y., and Hamamoto, R. Overexpression of the JmjC histone demethylase KDM5B in human carcinogenesis: involvement in the proliferation of cancer cells through the E2F/RB pathway. Mol Cancer, *9*: 59, 2010.
16. Takayama, R., Nakagawa, H., Sawaki, A., Mizuno, N., Kawai, H., Tajika, M., Yatabe, Y., Matsuo, K., Uehara, R., Ono, K., Nakamura, Y., and Yamao, K. Serum tumor antigen REG4 as a diagnostic biomarker in pancreatic ductal adenocarcinoma. J Gastroenterol, *45*: 52-59, 2010.
17. Yasuno, K., Bilguvar, K., Bijlenga, P., Low, S.K., Krischek, B., Auburger, G., Simon, M., Krex, D., Arlier, Z., Nayak, N., Ruigrok, Y. M., Niemela, M., Tajima, A., von und zu Fraunberg, M., Doczi, T., Wirjatijasa, F., Hata, A., Blasco, J., Oszvald, A., Kasuya, H., Zilani, G., Schoch, B., Singh, P., Stuer, C., Risselada, R., Beck, J., Sola, T., Ricciardi, F., Aromaa, A., Illig, T., Schreiber, S., van Duijn, C.M., van den Berg, L.H., Perret, C., Proust, C., Roder, C., Ozturk, A.K., Gaal, E., Berg, D., Geisen, C., Friedrich, C.M., Summers, P., Frangi, A.F., State, M.W., Wichmann, H.E., Breteler, M.M., Wijmenga, C., Mane, S., Peltonen, L., Elio, V., Sturkenboom, M.C., Lawford, P., Byrne, J., Macho, J., Sandalcioglu, E.I., Meyer, B., Raabe, A., Steinmetz, H., Rufenacht, D., Jaaskelainen, J. E., Hernesniemi, J., Rinkel, G.J., Zembutsu, H., Inoue, I., Palotie, A., Cambien, F., Nakamura, Y., Lifton, R.P., and Gunel, M. Genome-wide association study of intracranial aneurysm identifies three new risk loci. Nat Genet, *42*: 420-425, 2010.
18. Nakajima, M., Takahashi, A., Kou, I., Rodriguez-Fontenla, C., Gomez-Reino, J.J., Furuichi, T., Dai, J., Sudo, A., Uchida, A., Fukui, N., Kubo, M., Kamatani, N., Tsunoda, T., Malizos, K.N., Tsezou, A., Gonzalez, A., Nakamura, Y., and Ikegawa, S. New sequence variants in HLA class II/III region associated with susceptibility to knee osteoarthritis identified by genome-wide association study. PLoS One, *5*: e9723, 2010.
19. Hayami, S., Kelly, J.D., Cho, H.S., Yoshimatsu, M., Unoki, M., Tsunoda, T., Field, H. I., Neal, D.E., Yamaue, H., Ponder, B.A., Nakamura, Y., and Hamamoto, R. Overexpression of LSD1 contributes to human carcinogenesis through chromatin regulation in various cancers. Int J Cancer, *128*: 574-586, 2011.
20. Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabe, R.R., Bhan, M.K., Calvo, F., Eerola, I., Gerhard, D.S., Guttmacher, A., Guyer, M., Hemsley, F.M., Jennings, J.L., Kerr, D., Klatt, P., Kolar, P., Kusada, J., Lane, D.P., Laplace, F., Youyong, L., Nettekoven, G., Ozenberger, B., Peterson, J., Rao, T.S., Remacle, J., Schafer, A.J., Shibata, T., Stratton, M.R., Vockley, J.G., Watanabe, K., Yang, H., Yuen, M.M., Knoppers, B.M., Bobrow, M., Cambon-Thomsen, A., Dressler, L.G., Dyke, S.O., Joly, Y., Kato, K., Kennedy, K.L., Nicolas, P., Parker, M.J., Rial-Sebbag, E., Romeo-Casabona, C.M., Shaw, K. M., Wallace, S., Wiesner, G.L., Zeps, N., Lichter, P., Biankin, A.V., Chabannon, C., Chin, L., Clement, B., de Alava, E., Degos, F., Ferguson, M.L., Geary, P., Hayes, D.N., Hudson, T.J., Johns, A.L., Kasprzyk, A., Nakagawa, H., Penny, R., Piris, M.A., Sarin, R., Scarpa, A., Shibata, T., van de Vijver, M., Futreal, P.A., Aburatani, H., Bayes, M., Botwell, D.D., Campbell, P.J., Estivill, X., Gerhard, D.S., Grimmond, S.M., Gut, I., Hirst, M., Lopez-Otin, C., Majumder, P., Marra, M., McPherson, J.D., Nakagawa, H., Ning, Z., Puente, X.S., Ruan, Y., Shibata, T., Stratton, M.R., Stunnenberg, H.G., Swerdlow, H., Velculescu, V.E., Wilson, R.K., Xue, H.H., Yang, L., Spellman, P.T., Bader, G.D., Boutros, P.C., Campbell, P.J., Flicek, P., Getz, G., Guigo, R., Guo, G., Haussler, D., Heath, S., Hubbard, T.J., Jiang, T., Jones, S.M., Li, Q., Lopez-Bigas, N., Luo, R., Muthuswamy, L., Ouellette, B.F., Pearson, J.V., Puente, X.S., Quesada, V., Raphael, B.J., Sander, C., Shibata, T., Speed, T.P., Stein, L.D., Stuart, J.M., Teague, J.W., Totoki, Y., Tsunoda, T., Valencia, A., Wheeler, D.A., Wu, H., Zhao, S., Zhou, G., Stein, L.D., Guigo, R., Hubbard, T. J., Joly, Y., Jones, S.M., Kasprzyk, A., Lathrop, M., Lopez-Bigas, N., Ouellette, B.F., Spellman, P.T., Teague, J.W., Thomas, G., Valencia, A., Yoshida, T., Kennedy, K.L., Axton, M., Dyke, S.O., Futreal, P.A., Gerhard, D.S., Gunter, C., Guyer, M., Hudson, T.J., McPherson, J.D., Miller, L.J., Ozenberger, B., Shaw, K.M., Kasprzyk, A., Stein, L.D., Zhang, J., Haider, S.A., Wang, J., Yung, C.K., Cross, A., Liang, Y., Gnaneshan, S., Guberman, J., Hsu, J., Bobrow, M., Chalmers, D.R., Hasel, K.W., Joly, Y., Kaan, T.S., Kennedy, K.L., Knoppers, B.M., Lowrance, W.W., Masui, T., Nicolas, P., Rial-Sebbag, E., Rodriguez, L.L., Vergely, C., Yoshida, T., Grimmond, S.M., Biankin, A.V., Bowtell, D.D., Cloonan, N., deFazio, A., Eshleman, J.R., Etemadmoghadam, D., Gardiner, B.A., Kench, J.G., Scarpa, A., Sutherland, R.L., Tempero, M.A., Waddell, N.J., Wilson, P.J.,

McPherson, J.D., Gallinger, S., Tsao, M.S., Shaw, P.A., Petersen, G.M., Mukhopadhyay, D., Chin, L., DePinho, R.A., Thayer, S., Muthuswamy, L., Shazand, K., Beck, T., Sam, M., Timms, L., Ballin, V., Lu, Y., Ji, J., Zhang, X., Chen, F., Hu, X., Zhou, G., Yang, Q., Tian, G., Zhang, L., Xing, X., Li, X., Zhu, Z., Yu, Y., Yu, J., Yang, H., Lathrop, M., Tost, J., Brennan, P., Holcatova, I., Zaridze, D., Brazma, A., Egevard, L., Prokhortchouk, E., Banks, R.E., Uhlen, M., Cambon-Thomsen, A., Viksna, J., Ponten, F., Skryabin, K., Stratton, M.R., Futreal, P.A., Birney, E., Borg, A., Borresen-Dale, A.L., Caldas, C., Foekens, J.A., Martin, S., Reis-Filho, J.S., Richardson, A.L., Sotiriou, C., Stunnenberg, H.G., Thoms, G., van de Vijver, M., van't Veer, L., Calvo, F., Birnbaum, D., Blanche, H., Boucher, P., Boyault, S., Chabannon, C., Gut, I., Masson-Jacquemier, J.D., Lathrop, M., Pauporte, I., Pivot, X., Vincent-Salomon, A., Tabone, E., Theillet, C., Thomas, G., Tost, J., Treilleux, I., Calvo, F., Bioulac-Sage, P., Clement, B., Decaens, T., Degos, F., Franco, D., Gut, I., Gut, M., Heath, S., Lathrop, M., Samuel, D., Thomas, G., Zucman-Rossi, J., Lichter, P., Eils, R., Brors, B., Korbel, J.O., Korshunov, A., Landgraf, P., Lehrach, H., Pfister, S., Radlwimmer, B., Reifenberger, G., Taylor, M.D., von Kalle, C., Majumder, P.P., Sarin, R., Rao, T.S., Bhan, M.K., Scarpa, A., Pederzoli, P., Lawlor, R.A., Delledonne, M., Bardelli, A., Biankin, A.V., Grimmond, S.M., Gress, T., Klimstra, D., Zamboni, G., Shibata, T., Nakamura, Y., Nakagawa, H., Kusada, J., Tsunoda, T., Miyano, S., Aburatani, H., Kato, K., Fujimoto, A., Yoshida, T., Campo, E., Lopez-Otin, C., Estivill, X., Guigo, R., de Sanjose, S., Piris, M.A., Montserrat, E., Gonzalez-Diaz, M., Puente, X.S., Jares, P., Valencia, A., Himmelbaue, H., Quesada, V., Bea, S., Stratton, M.R., Futreal, P.A., Campbell, P.J., Vincent-Salomon, A., Richardson, A.L., Reis-Filho, J.S., van de Vijver, M., Thomas, G., Masson-Jacquemier, J.D., Aparicio, S., Borg, A., Borresen-Dale, A.L., Caldas, C., Foekens, J.A., Stunnenberg, H.G., van't Veer, L., Easton, D.F., Spellman, P.T., Martin, S., Barker, A.D., Chin, L., Collins, F.S., Compton, C.C., Ferguson, M.L., Gerhard, D.S., Getz, G., Gunter, C., Guttmacher, A., Guyer, M., Hayes, D.N., Lander, E.S., Ozenberger, B., Penny, R., Peterson, J., Sander, C., Shaw, K.M., Speed, T.P., Spellman, P.T., Vockley, J. G., Wheeler, D.A., Wilson, R.K., Hudson, T. J., Chin, L., Knoppers, B.M., Lander, E.S., Lichter, P., Stein, L.D., Stratton, M.R., Anderson, W., Barker, A.D., Bell, C., Bobrow, M., Burke, W., Collins, F.S., Compton, C.C., DePinho, R.A., Easton, D.F., Futreal, P.A., Gerhard, D.S., Green, A.R., Guyer, M., Hamilton, S.R., Hubbard, T.J., Kallioniemi, O.P., Kennedy, K.L., Ley, T.J., Liu, E.T., Lu, Y., Majumder, P., Marra, M., Ozenberger, B., Peterson, J., Schafer, A.J., Spellman, P.T., Stunnenberg, H.G., Wainwright, B.J., Wilson, R.K. and Yang, H. International network of cancer genome projects. Nature, *464*: 993-998, 2010.

21. Yamaguchi, K., Sakai, M., Shimokawa, T., Yamada, Y., Nakamura, Y., and Furukawa, Y. C20orf20 (MRG-binding protein) as a potential therapeutic target for colorectal cancer. Br J Cancer, *102*: 325-331, 2010.
22. Kashiwaya, K., Nakagawa, H., Hosokawa, M., Mochizuki, Y., Ueda, K., Piao, L., Chung, S., Hamamoto, R., Eguchi, H., Ohigashi, H., Ishikawa, O., Janke, C., Shinomura, Y., and Nakamura, Y. Involvement of the tubulin tyrosine ligase-like family member 4 polyglutamylase in PELP1 polyglutamylation and chromatin remodeling in pancreatic cancer cells. Cancer Res, *70*: 4024-4033, 2010.
23. Akuta, N., Suzuki, F., Hirakawa, M., Kawamura, Y., Yatsuji, H., Sezaki, H., Suzuki, Y., Hosaka, T., Kobayashi, M., Kobayashi, M., Saitoh, S., Arase, Y., Ikeda, K., Chayama, K., Nakamura, Y., and Kumada, H. Amino acid substitution in hepatitis C virus core region and genetic variation near the interleukin 28 B gene predict viral response to telaprevir with peginterferon and ribavirin. Hepatology, *52*: 421-429, 2010.
24. Yoshimatsu, M., Toyokawa, G., Hayami, S., Unoki, M., Tsunoda, T., Field, H.I., Kelly, J. D., Neal, D.E., Maehara, Y., Ponder, B.A., Nakamura, Y., and Hamamoto, R. Dysregulation of PRMT1 and PRMT6, Type I arginine methyltransferases, is involved in various types of human cancers. Int J Cancer, *128*: 562-573, 2011.
25. Phasukkijwatana, N., Kunhapan, B., Stankovich, J., Chuenkongkaew, W.L., Thomson, R., Thornton, T., Bahlo, M., Mushiroda, T., Nakamura, Y., Mahasirimongkol, S., Tun, A. W., Srisawat, C., Limwongse, C., Peerapittayamongkol, C., Sura, T., Suthammarak, W., and Lertrit, P. Genome-wide linkage scan and association study of PARL to the expression of LHON families in Thailand. Hum Genet, *128*: 39-49, 2010.
26. Ramamoorthy, A., Flockhart, D.A., Hosono, N., Kubo, M., Nakamura, Y., and Skaar, T.C. Differential quantification of CYP2D6 gene copy number by four different quantitative real-time PCR assays. Pharmacogenet Genomics, *20*: 451-454, 2010.

27. Kochi, Y., Okada, Y., Suzuki, A., Ikari, K., Terao, C., Takahashi, A., Yamazaki, K., Hosono, N., Myouzen, K., Tsunoda, T., Kamatani, N., Furuichi, T., Ikegawa, S., Ohmura, K., Mimori, T., Matsuda, F., Iwamoto, T., Momohara, S., Yamanaka, H., Yamada, R., Kubo, M., Nakamura, Y., and Yamamoto, K. A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. Nat Genet, *42*: 515-519, 2010.
28. Kiyotani, K., Mushiroda, T., Hosono, N., Tsunoda, T., Kubo, M., Aki, F., Okazaki, Y., Hirata, K., Takatsuka, Y., Okazaki, M., Ohsumi, S., Yamakawa, T., Sasa, M., Nakamura, Y., and Zembutsu, H. Lessons for pharmacogenomics studies : association study between CYP2D6 genotype and tamoxifen response. Pharmacogenet Genomics, *20*: 565-568, 2010.
29. Abe, H., Ochi, H., Maekawa, T., Hayes, C. N., Tsuge, M., Miki, D., Mitsui, F., Hiraga, N., Imamura, M., Takahashi, S., Ohishi, W., Arihiro, K., Kubo, M., Nakamura, Y., and Chayama, K. Common variation of IL28 affects gamma-GTP levels and inflammation of the liver in chronically infected hepatitis C virus patients. J Hepatol, *53*: 439-443, 2010.
30. Tanikawa, C., Ri, C., Kumar, V., Nakamura, Y., and Matsuda, K. Crosstalk of EDA-A2/XEDAR in the p53 signaling pathway. Mol Cancer Res, *8*: 855-863, 2010.
31. Onouchi, Y., Ozaki, K., Buns, J.C., Shimizu, C., Hamada, H., Honda, T., Terai, M., Honda, A., Takeuchi, T., Shibuta, S., Suenaga, T., Suzuki, H., Higashi, K., Yasukawa, K., Suzuki, Y., Sasago, K., Kemmotsu, Y., Takatsuki, S., Saji, T., Yoshikawa, T., Nagai, T., Hamamoto, K., Kishi, F., Ouchi, K., Sato, Y., Newburger, J.W., Baker, A.L., Shulman, S.T., Rowley, A.H., Yashiro, M., Nakamura, Y., Wakui, K., Fukushima, Y., Fujino, A., Tsunoda, T., Kawasaki, T., Hata, A., Nakamura, Y., and Tanaka, T. Common variants in CASP3 confer susceptibility to Kawasaki disease. Hum Mol Genet, *19*: 2898-2906, 2010.
32. Unoki, M., Kelly, J.D., Neal, D.E., Ponder, B. A., Nakamura, Y., and Hamamoto, R. UHRF 1 is a novel molecular marker for diagnosis and the prognosis of bladder cancer. Br J Cancer, *101*: 98-105, 2009.
33. Nguyen, M.H., Koinuma, J., Ueda, K., Ito, T., Tsuchiya, E., Nakamura, Y., and Daigo, Y. Phosphorylation and activation of cell division cycle associated 5 by mitogen-activated protein kinase play a crucial role in human lung carcinogenesis. Cancer Res, *70*: 5337-5347, 2010.
34. Fukukawa, C., Ueda, K., Nishidate, T., Katagiri, T., and Nakamura, Y. Critical roles of LGN/GPSM2 phosphorylation by PBK/TOPK in cell division of breast cancer cells. Genes Chromosomes Cancer, *49*: 861-872, 2010.
35. Sato, N., Yamabuki, T., Takano, A., Koinuma, J., Aragaki, M., Masuda, K., Ishikawa, N., Kohno, N., Ito, H., Miyamoto, M., Nakayama, H., Miyagi, Y., Tsuchiya, E., Kondo, S., Nakamura, Y., and Daigo, Y. Wnt inhibitor Dickkopf-1 as a target for passive cancer immunotherapy. Cancer Res, *70*: 5326-5336, 2010.
36. Harada, Y., Kanehira, M., Fujisawa, Y., Takata, R., Shuin, T., Miki, T., Fujioka, T., Nakamura, Y., and Katagiri, T. Cell-permeable peptide DEPDC1-ZNF224 interferes with transcriptional repression and oncogenicity in bladder cancer cells. Cancer Res, *70*: 5829-5839, 2010.
37. Uno, S., Zembutsu, H., Hirasawa, A., Takahashi, A., Kubo, M., Akahane, T., Aoki, D., Kamatani, N., Hirata, K., and Nakamura, Y. A genome-wide association study identifies genetic variants in the CDKN2BAS locus associated with endometriosis in Japanese. Nat Genet, *42*: 707-710, 2010.
38. Uemura, M., Honma, S., Chung, S., Takata, R., Furihata, M., Nishimura, K., Nonomura, N., Nasu, Y., Miki, T., Shuin, T., Fujioka, T., Okuyama, A., Nakamura, Y., and Nakagawa, H. 5alphaDH-DOC (5alpha-dihydrodeoxycorticosterone) activates androgen receptor in castration-resistant prostate cancer. Cancer Sci, *101*: 1897-1904, 2010.
39. Akaza, H., Kawai, K., Tsukamoto, T., Fujioka, T., Tomita, Y., Kitamura, T., Ozono, S., Miki, T., Naito, S., Zembutsu, H., and Nakamura, Y. Successful outcomes using combination therapy of interleukin-2 and interferon-alpha for renal cell carcinoma patients with lung metastasis. Jpn J Clin Oncol, *40*: 684-689, 2010.
40. Ueda, K., Takami, S., Saichi, N., Daigo, Y., Ishikawa, N., Kohno, N., Katsumata, M., Yamane, A., Ota, M., Sato, T.A., Nakamura, Y., and Nakagawa, H. Development of serum glycoproteomic profiling technique; simultaneous identification of glycosylation sites and site-specific quantification of glycan structure changes. Mol Cell Proteomics, *9*: 1819-1828, 2010.
41. Ingle, J.N., Schaid, D.J., Goss, P.E., Liu, M., Mushiroda, T., Chapman, J.A., Kubo, M., Jenkins, G.D., Batzler, A., Shepherd, L., Pater, J., Wang, L., Ellis, M.J., Stearns, V., Rohrer, D.C., Goetz, M.P., Pritchard, K.I., Flockhart, D.A., Nakamura, Y., and Wein-

shilboum, R.M. Genome-wide associations and functional genomic studies of musculoskeletal adverse events in women receiving aromatase inhibitors. J Clin Oncol, *28*: 4674-4682, 2010.
42. Okada, Y., Kamatani, Y., Takahashi, A., Matsuda, K., Hosono, N., Ohmiya, H., Daigo, Y., Yamamoto, K., Kubo, M., Nakamura, Y., and Kamatani, N. Common variations in PSMD3-CSF3 and PLCB4 are associated with neutrophil count. Hum Mol Genet, *19*: 2079-2085, 2010.
43. Okada, Y., Kamatani, Y., Takahashi, A., Matsuda, K., Hosono, N., Ohmiya, H., Daigo, Y., Yamamoto, K., Kubo, M., Nakamura, Y., and Kamatani, N. A genome-wide association study in 19 633 Japanese subjects identified LHX3-QSOX2 and IGF1 as adult height loci. Hum Mol Genet, *19*: 2303-2312, 2010.
44. Takata, R., Akamatsu, S., Kubo, M., Takahashi, A., Hosono, N., Kawaguchi, T., Tsunoda, T., Inazawa, J., Kamatani, N., Ogawa, O., Fujioka, T., Nakamura, Y., and Nakagawa, H. Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. Nat Genet, *42*: 751-754, 2010.
45. Kim, J.W., Fukukawa, C., Ueda, K., Nishidate, T., Katagiri, T., and Nakamura, Y. Involvement of C12orf32 overexpression in breast carcinogenesis. Int J Oncol, *37*: 861-867, 2010.
46. Nakashima, M., Chung, S., Takahashi, A., Kamatani, N., Kawaguchi, T., Tsunoda, T., Hosono, N., Kubo, M., Nakamura, Y., and Zembutsu, H. A genome-wide association study identifies four susceptibility loci for keloid in the Japanese population. Nat Genet, *42*: 768-771, 2010.
47. Iida, A., Kamei, T., Sano, M., Oshima, S., Tokuda, T., Nakamura, Y., and Ikegawa, S. Large-scale screening of TARDBP mutation in amyotrophic lateral sclerosis in Japanese. Neurobiol Aging, 2010.
48. Low, S.K., Kuchiba, A., Zembutsu, H., Saito, A., Takahashi, A., Kubo, M., Daigo, Y., Kamatani, N., Chiku, S., Totsuka, H., Ohnami, S., Hirose, H., Shimada, K., Okusaka, T., Yoshida, T., Nakamura, Y., and Sakamoto, H. Genome-wide association study of pancreatic cancer in Japanese population. PLoS One, *5*: e11824, 2010.
49. Kawaoka, T., Hayes, C. N., Ohishi, W., Ochi, H., Maekawa, T., Abe, H., Tsuge, M., Mitsui, F., Hiraga, N., Imamura, M., Takahashi, S., Kubo, M., Tsunoda, T., Nakamura, Y., Kumada, H., and Chayama, K. Predictive value of the IL28B polymorphism on the effect of interferon therapy in chronic hepatitis C patients with genotypes 2a and 2b. J Hepatol, 2010.
50. Ochi, H., Maekawa, T., Abe, H., Hayashida, Y., Nakano, R., Kubo, M., Tsunoda, T., Hayes, C.N., Kumada, H., Nakamura, Y., and Chayama, K. ITPA polymorphism affects ribavirin-induced anemia and outcomes of therapy--a genome-wide study of Japanese HCV virus patients. Gastroenterology, *139*: 1190-1197, 2010.
51. Ajiro, M., Nishidate, T., Katagiri, T., and Nakamura, Y. Critical involvement of RQCD 1 in the EGFR-Akt pathway in mammary carcinogenesis. Int J Oncol, *37*: 1085-1093, 2010.
52. Miki, D., Kubo, M., Takahashi, A., Yoon, K. A., Kim, J., Lee, G.K., Zo, J.I., Lee, J.S., Hosono, N., Morizono, T., Tsunoda, T., Kamatani, N., Chayama, K., Takahashi, T., Inazawa, J., Nakamura, Y., and Daigo, Y. Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations. Nat Genet, *42*: 893-896, 2010.
53. Myouzen, K., Kochi, Y., Shimane, K., Fujio, K., Okamura, T., Okada, Y., Suzuki, A., Atsumi, T., Ito, S., Takada, K., Mimori, A., Ikegawa, S., Yamada, R., Nakamura, Y., and Yamamoto, K. Regulatory polymorphisms in EGR2 are associated with susceptibility to systemic lupus erythematosus. Hum Mol Genet, *19*: 2313-2320, 2010.
54. Yamauchi, T., Hara, K., Maeda, S., Yasuda, K., Takahashi, A., Horikoshi, M., Nakamura, M., Fujita, H., Grarup, N., Cauchi, S., Ng, D. P., Ma, R.C., Tsunoda, T., Kubo, M., Watada, H., Maegawa, H., Okada-Iwabu, M., Iwabu, M., Shojima, N., Shin, H.D., Andersen, G., Witte, D.R., Jorgensen, T., Lauritzen, T., Sandbaek, A., Hansen, T., Ohshige, T., Omori, S., Saito, I., Kaku, K., Hirose, H., So, W.Y., Beury, D., Chan, J.C., Park, K.S., Tai, E.S., Ito, C., Tanaka, Y., Kashiwagi, A., Kawamori, R., Kasuga, M., Froguel, P., Pedersen, O., Kamatani, N., Nakamura, Y., and Kadowaki, T. A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE 2E2 and C2CD4A-C2CD4B. Nat Genet, *42*: 864-868, 2010.
55. Akamatsu, S., Takata, R., Ashikawa, K., Hosono, N., Kamatani, N., Fujioka, T., Ogawa, O., Kubo, M., Nakamura, Y., and Nakagawa, H. A functional variant in NKX 3.1 associated with prostate cancer susceptibility down-regulates NKX3.1 expression. Hum Mol Genet, *19*: 4265-4272, 2010.
56. Okada, Y., Suzuki, A., Yamada, R., Kochi,

Y., Shimane, K., Myouzen, K., Kubo, M., Nakamura, Y., and Yamamoto, K. HLA-DRB 1*0901 lowers anti-cyclic citrullinated peptide antibody levels in Japanese patients with rheumatoid arthritis. Ann Rheum Dis, *69*: 1569-1570, 2010.

57. Cha, P.C., Mushiroda, T., Takahashi, A., Kubo, M., Minami, S., Kamatani, N., and Nakamura, Y. Genome-wide association study identifies genetic determinants of warfarin responsiveness for Japanese. Hum Mol Genet, *19*: 4735-4744, 2010.
58. Chung, S., Nakagawa, H., Uemura, M., Piao, L., Ashikawa, K., Hosono, N., Takata, R., Akamatsu, S., Kawaguchi, T., Morizono, T., Tsunoda, T., Daigo, Y., Matsuda, K., Kamatani, N., Nakamura, Y., and Kubo, M. Association of a novel long non-coding RNA in 8 q24 with prostate cancer susceptibility. Cancer Sci, *102*: 245-252, 2011.
59. Kumasaka, N., Yamaguchi-Kabata, Y., Takahashi, A., Kubo, M., Nakamura, Y., and Kamatani, N. Establishment of a standardized system to perform population structure analyses with limited sample size or with different sets of SNP genotypes. J Hum Genet, *55*: 525-533, 2010.
60. Piao, L., Nakagawa, H., Ueda, K., Chung, S., Kashiwaya, K., Eguchi, H., Ohigashi, H., Ishikawa, O., Daigo, Y., Matsuda, K., and Nakamura, Y. C12orf48, termed PARP-1 binding protein, enhances poly(ADP-ribose) polymerase-1 (PARP-1) activity and protects pancreatic cancer cells from DNA damage. Genes Chromosomes Cancer, *50*: 13-24, 2011.
61. Stefanou, N., Papanikolaou, V., Furukawa, Y., Nakamura, Y., and Tsezou, A. Leptin as a critical regulator of hepatocellular carcinoma development through modulation of human telomerase reverse transcriptase. BMC Cancer, *10*: 442, 2010.
62. Yamaguchi-Kabata, Y., Tsunoda, T., Takahashi, A., Hosono, N., Kubo, M., Nakamura, Y., and Kamatani, N. Making a haplotype catalog with estimated frequencies based on SNP homozygotes. J Hum Genet, *55*: 500-506, 2010.
63. Hayes, C.N., Kobayashi, M., Akuta, N., Suzuki, F., Kumada, H., Abe, H., Miki, D., Imamura, M., Ochi, H., Kamatani, N., Nakamura, Y., and Chayama, K. HCV substitutions and IL28B polymorphisms on outcome of peg-interferon plus ribavirin combination therapy. Gut, *60*: 261-267, 2011.
64. Fujimoto, A., Nakagawa, H., Hosono, N., Nakano, K., Abe, T., Boroevich, K.A., Nagasaki, M., Yamaguchi, R., Shibuya, T., Kubo, M., Miyano, S., Nakamura, Y., and Tsunoda, T. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nat Genet, *42*: 931-936, 2010.
65. Aoki, A., Ozaki, K., Sato, H., Takahashi, A., Kubo, M., Sakata, Y., Onouchi, Y., Kawaguchi, T., Lin, T.H., Takano, H., Yasutake, M., Hsu, P.C., Ikegawa, S., Kamatani, N., Tsunoda, T., Juo, S.H., Hori, M., Komuro, I., Mizuno, K., Nakamura, Y., and Tanaka, T. SNPs on chromosome 5p15.3 associated with myocardial infarction in Japanese population. J Hum Genet, 2010.
66. Fujimoto, Y., Ochi, H., Maekawa, T., Abe, H., Hayes, C.N., Kumada, H., Nakamura, Y., and Chayama, K. A single nucleotide polymorphism in activated cdc42 associated tyrosine kinase 1 influences the interferon therapy in hepatitis C patients. J Hepatol, 2010.
67. Cho, H.S., Suzuki, T., Dohmae, N., Hayami, S., Unoki, M., Yoshimatsu, M., Toyokawa, G., Takawa, M., Chen, T., Kurash, J.K., Field, H., Ponder, B.A., Nakamura, Y., and Hamamoto, R. Demethylation of RB regulator MYPT1 by histone demethylase LSD1 promotes cell cycle progression in cancer cells. Cancer Res, 2010.
68. Maeda, S., Araki, S., Babazono, T., Toyoda, M., Umezono, T., Kawai, K., Imanishi, M., Uzu, T., Watada, H., Suzuki, D., Kashiwagi, A., Iwamoto, Y., Kaku, K., Kawamori, R., and Nakamura, Y. Replication study for the association between four Loci identified by a genome-wide association study on European American subjects with type 1 diabetes and susceptibility to diabetic nephropathy in Japanese subjects with type 2 diabetes. Diabetes, *59*: 2075-2079, 2010.
69. Iida, A., Hosono, N., Sano, M., Kamei, T., Oshima, S., Tokuda, T., Kubo, M., Nakamura, Y., and Ikegawa, S. Optineurin mutations in Japanese amyotrophic lateral sclerosis. J Neurol Neurosurg Psychiatry, 2011.
70. Saetre, P., Vares, M., Werge, T., Andreassen, O.A., Arinami, T., Ishiguro, H., Nanko, S., Tan, E.C., Han, D.H., Roffman, J.L., Muntjewerff, J.W., Jagodzinski, P.P., Kempisty, B., Hauser, J., Vilella, E., Betcheva, E., Nakamura, Y., Regland, B., Agartz, I., Hall, H., Terenius, L., and Jonsson, E.G. Methylenetetrahydrofolate reductase (MTHFR) C677T and A1298C polymorphisms and age of onset in schizophrenia: A combined analysis of independent samples. Am J Med Genet B Neuropsychiatr Genet, 2011.
71. Ozeki, T., Mushiroda, T., Yowang, A., Takahashi, A., Kubo, M., Shirakata, Y., Ikezawa,

Z., Iijima, M., Shiohara, T., Hashimoto, K., Kamatani, N., and Nakamura, Y. Genome-wide association study identifies HLA-A* 3101 allele as a genetic risk factor for carbamazepine-induced cutaneous adverse drug reactions in Japanese population. Hum Mol Genet, 2010.

72. Yoon, K.A., Park, J.H., Han, J., Park, S., Lee, G.K., Han, J.Y., Zo, J.I., Kim, J., Lee, J.E., Takahashi, A., Kubo, M., Nakamura, Y., and Lee, J.S. A genome-wide association study reveals susceptibility variants for non-small cell lung cancer in the Korean population. Hum Mol Genet, *19*: 4948-4954, 2010.
73. Peerbooms, O.L., van Os, J., Drukker, M., Kenis, G., Hoogveld, L., de Hert, M., Delespaul, P., van Winkel, R., and Rutten, B. P. Meta-analysis of MTHFR gene variants in schizophrenia, bipolar disorder and unipolar depressive disorder: Evidence for a common genetic vulnerability? Brain Behav Immun, 2010.
74. Imai, K., Hirata, S., Irie, A., Senju, S., Ikuta, Y., Yokomine, K., Harao, M., Inoue, M., Tomita, Y., Tsunoda, T., Nakagawa, H., Nakamura, Y., Baba, H., and Nishimura, Y. Identification of HLA-A2-restricted CTL epitopes of a novel tumour-associated antigen, KIF20A, overexpressed in pancreatic cancer. Br J Cancer, 2010.
75. Mizumori, O., H. Zembutsu, Y. Kato, T. Tsunoda, F. Miya, T. Morizono, T. Tsukamoto, T. Fujioka, Y. Tomita, T. Kitamura, S. Ozono, T. Miki, S. Naito, H. Akaza, and Y. Nakamura. Identification of a set of genes associated with response to interleukin-2 and interferon-α combination therapy for renal cell carcinoma through genome-wide gene expression profiling. Experimental and Therapeutic Medicine. *1*, 955-961.2010.
76. Kato. Y., H. Zembutsu, R. Takata, F. Miya, T. Tsunoda, W. Obara, T. Fujioka, and Y. Nakamura: Predicting response of bladder cancers to gemcitabine and carboplatin neoadjuvant chemotherapy through genome -wide gene expression profiling. Experimental and Therapeutic Medicine In press.
77. Hashimoto, Y., H. Ochi, H. Abe, Y. Hayashida, M. Tsuge, F. Mitsui, N. Hiraga, M. Imamura, S. Takahashi, C.N. Hayes, W. Ohishi, M. Kubo, T. Tsunoda, N. Kamatani, Y. Nakamura, and K. Chayama: Prediction of response to peginterferon-alfa-2b plus ribavirin therapy in Japanese patients infected with hepatitis C virus genotype 1b. J. Med. Virol. In press
78. Suzuki, F., Y. Suzuki, N. Akuta, H. Sezaki, M. Hirakawa, Y. Kawamura, T. Hosaka, M. Kobayashi, S. Saito, Y. Arase, K. Ikeda, M. Kobayashi, K. Chayama, N. Kamatani, Y. Nakamura, Y. Miyakawa, and H. Kumada: Influence of ITPA polymorphism on decreases of hemoglobin during treatment with pegylated IFN, ribavirin and telaprevir. Hepatology In press
79. Cui, R., Y. Okada, S.G. Jang, J.L. Ku, J.G. Park, Y. Kamatani, N. Hosono, T. Tsunoda, V. Kumar, C. Tanikawa, N. Kamatani, R. Yamada, M. Kubo, Y. Nakamura, and K. Matsuda: Common variant in 6q26-q27 is associated with distal colon cancer in Asian population. GUT In press.
80. Akukta, N., F. Suzuki, M. Hirakawa, Y. Kawamura, H. Yatsuji, H. Sezaki, Y. Suzuki, T. Hosaka, M. Kobayashi, M. Kobayashi, S. Saitoh, Y. Arase, K. Ikeda, K. Chayama, Y. Nakamura, and H. Kumada. Amino acid substitution in HCV core region and genetic variation near IL28B gene affect viral dynamics during telaprevir, peginterferon and ribavirin. Intervirology In press.

Human Genome Center

Laboratory of Functional Analysis *In Silico*
機能解析イン・シリコ分野

Professor	Kenta Nakai, Ph.D.	教　授　理学博士	中　井　謙　太
Assistant Professor	Ashwini Patil, Ph.D.	助　教　理学博士	パティル，アシュウィニ

The mission of our laboratory is to conduct computational ("in silico") studies on the functional aspects of genome information. Roughly speaking, genome information represents what kind of proteins/RNAs are synthesized in what conditions. Thus, our study includes the structural analysis of molecular function of each gene product as well as the analysis of its regulatory information, which will lead us to the understanding of its cellular role represented by the networks of inter-gene interaction.

1. A regression analysis of gene expression in ES cells reveals two gene classes that are significantly different in epigenetic patterns

Sung-Joon Park and Kenta Nakai

To understand the gene regulatory system that governs the pluripotency of embryonic stem (ES) cells is an important step for promoting regenerative medicine. Although the role of several transcription factors (TFs) has been intensively investigated, details of their involvement in the regulation are still not well clarified. Here we constructed a predictive model of the genome-wide gene expression in mouse ES cells. We firstly reanalyzed ChIP-seq data publicly available. Then, we estimated TF-binding density profiles from the data. The density profiles and the data of several epigenetic states of promoters are used as predictors in a simple linear regression model that predicts absolute gene expression. We also exhaustively analyzed the effects of predictors and their higher-order interactions by statistical tests. Through this study, we confirmed that our linear regression model has better predictive power than an ordinary linear regression model. Using the proposed model, we identified two gene classes that are either well explained or inefficiently explained by the model. Since the promoters of these gene classes have apparently distinct patterns of epigenetics, we inspected the effects of epigenetic modifications in explaining the gene expression. The result supports the general idea of relative importance of epigenetic effects in ES cells. Moreover, it has been found that unknown regulatory mechanisms control approximately half of ES-specific genes rather than direct regulation by the TFs. To elucidate such mechanisms is one of further works.

2. Transcriptional regulation and networks of immune systems

Naoki Osato and Kenta Nakai

Our immune responses involve highly organized activities of many types of genes and cells. Understanding the mechanisms of immune responses is important to reveal the causes of diseases, allergies, and immunological rejections. Although these mechanisms have been examined by experimental analyses, they are still unclear. Nowadays, genome-wide gene expression data, protein-protein and protein-DNA interac-

tion data, and epigenetic data are being produced. Analyzing genome sequences and these data would contribute to reveal the mechanisms and regulatory networks of immune responses and the differentiation of immune cells effectively. For these purposes, we plan to develop methods to predict transcription factor binding sites (TFBS) from human genome sequences and gene expression data, and reveal conserved signatures of TFBS in each immune cell type or response by computational analyses. We will predict transcriptional regulatory networks of immune cells and responses by combining various genome-scale data such as known TFBS, noncoding RNAs, and known pathways. Collaborating with experimental researchers, we will validate our predictions. These methods and approaches would also be useful to examine other transcriptional regulatory mechanisms and networks.

3. Classification and characterization of bidirectional promoters.

Riu Yamashita[1], Yutaka Suzuki[2], Sumio Sugano[2], Kenta Nakai: [1]Frontier Research Initiative, IMSUT, [2]Graduate School of Frontier Sciences

We have constructed the DataBase of Transcription Start Sites (DBTSS: http://dbtss.hgc.jp/), which contains the information of accurate transcription start sites (TSSs) based on experimentally determined 5′-end clones. Recently we have updated the database based on the 5′ end of oligo-cap selected cDNAs in humans and mice. Using this database, we focused on bidirectional promoters, which have TSSs on both plus and minus strands closely. In DLD-1 cell lines, the distribution of frequency of the distance between a transcript's TSS and the closest anti-transcript TSS showed two significant peaks. The first peak corresponded to 707 TSSs on −160:−100 (upstream anti-sense transcript promoters: upASTPs), and the second corresponded to 237 TSSs 0:+40 (downstream anti-sense transcript promoters: downASTPs). We also defined 2166 TSSs which did not have any anti-transcript within 10 kb as no anti-sense transcript promoters (noASTPs). Around TSS region, noASTPs and upASTPs had pyrimidine-purine bases on −1:+1, which can be widely observed on TSSs. However, the downASTPs showed 'CTGG' at −4:−1 region. We also observed significant difference of characteristics in the downASTPs: namely, GC poor, CpG islands poor, and disordered nucleosome structures. Interestingly, even though we observed highly ordered nucleosome structures in both downASTPs and noASTPs, the noASTPs showed more asymmetrical nucleosome structures. These phenomena could be observed not only in other human cell lines (Hek293, MCF7, TIG), but also in a mouse cell line (3T3). These results indicate that we could classify promoters into three classes based on their anti-transcript, and these classes showed biologically different features.

4. The effect of Alu elements on global nucleosome positioning in the human genome using paired-end MNase-Seq

Yoshiaki Tanaka, Riu Yamashita[1], Yutaka Suzuki[2] and Kenta Nakai

Understanding the genome sequence-specific positioning of nucleosomes is essential to understand various cellular processes, such as transcriptional regulation and replication. As a typical example, the 10-bp periodicity of AA/TT and GC dinucleotides has been reported in several species, but it is still unclear whether this feature can be observed in the whole genomes of all eukaryotes. With Fourier analysis, we found that this is not the case: 84-bp and 167-bp periodicities are prevalent in primates. The 167-bp periodicity is intriguing because it is almost equal to the sum of the lengths of a nucleosomal unit and its linker region. After masking Alu elements, these periodicities were greatly diminished. Next, using paired-end sequencing of nucleosomal DNAs (MNase-Seq), we analyzed the distribution of nucleosomes in the vicinity of Alu elements and showed that (1) there are one or two fixed slot(s) for nucleosome positioning within the Alu element and (2) the positioning of neighboring nucleosomes seems to be in phase, more or less, with the presence of Alu elements. The paired-end data makes it possible to reduce the number of multiple hits. Furthermore, (3) these effects of Alu elements on nucleosome positioning are consistent with inactivation of promoter activity in Alu elements. Our discoveries suggest that the principle governing nucleosome positioning differs greatly across species and that the Alu family is an important factor in primate genomes.

5. Prediction of subcellular locations of proteins: where to proceed?

Kenichiro Imai[3] and Kenta Nakai: [3]Computational Biology Research Center, AIST

Since the proposal of the signal hypothesis on protein subcellular sorting, a number of computational analyses have been performed in this

field. A typical example is the development of prediction algorithms for the subcellular localization sites of input protein sequences. In this review, we mainly focus on the biological grounds of the prediction methods rather than the algorithmic issues because we believe the former will be more fruitful for future development. Recent advances on the study of protein sorting signals will hopefully be incorporated into future prediction methods. Unfortunately, many of the state-of-the-art methods are published without sufficiently objective tests. In fact, a simple test employed in this article shows that the performance of specifically developed predictors is not significantly better than that of a homology search. We suspect this is a general problem associated with the interpretation of genome sequences, which have evolved through gene duplication and speciation.

6. Computational prediction of mitochondrial inner membrane proteins

Toshiyuki Tsuji and Kenta Nakai

Proteins are transferred to the correct subcellular location to properly perform their functions. Prediction of the protein subcellular localization is very important for exploring the function of proteins in the cell. Although many subcellular localization prediction methods have been developed, these systems ignore the mitochondrial inner membrane proteins without N-terminal targeting signal. It is known that the inner membrane proteins have an internal transit noncleavable signal. However the detection of the internal signals is very difficult. In this study, we analyzed the propensities of amino acids and various physicochemical properties of five types of proteins, which are localized at the mitochondrial inner membrane, the mitochondrial matrix, cytoplasm, the cell membrane or the extracellular space. These properties were used for developing a prediction method based on support vector machine (SVM). We analyzed the amino acid propensities around the cleavage site of the mitochondria targeting signal at the N-terminus of mitochondrial matrix proteins and made a position specific score matrix (PSSM). Using the PSSM, we searched the potential cleavage site of mitochondrial inner membrane proteins. The distributions of physicochemical properties around the potential cleavage site were used as the input of SVM. Meanwhile, we revealed that the mitochondrial inner membrane proteins tend to dislike the cluster of negative charged residues. We also used the size of negatively charged cluster as an input of SVM. The performance of our predictor outperforms previous predictors in the discrimination of the mitochondrial inner membrane proteins.

7. Development and maintenance of the databases Hintdb and HitPredict - Ashwini Patil, Kenta Nakai and Haruki Nakamura[4]: [4]Osaka University

Protein-protein interactions (PPIs) are vital for cellular function in organisms and hence their detection is of considerable importance. The advent of high-throughput technologies has lead to a manifold increase in the PPI information in several model organisms through large scale yeast two hybrid (Y2H) and tandem affinity purification in combination with mass spectrometry (TAP/MS) experiments. However, this data has two major drawbacks leading to its limited usage-(i) the large number of spurious interactions detected and (ii) the absence of direct binary interaction information in protein co-complex data obtained from TAP/MS experiments. As a result, most studies using PPI information either use data obtained exclusively from small-scale experiments, or those confirmed in multiple experiments. Both types of interaction subsets are considered high confidence but constitute only a fraction of the amount of data available and their use can often lead to biased results. An alternative approach is to utilize the high confidence subsets provided by authors of high-throughput experiments. However, these interaction subsets are assessed using a range of techniques with differing accuracies making comparisons among data sets difficult. Frequently, such high confidence interaction subsets are available only for one or two species, typically yeast and human. As a result, a large amount of the PPI information in several species, though correct and potentially useful, is often ignored. The major reason for this lack of information usage is the scarcity of comprehensive PPI databases that provide confidence scores assessing the quality of the interactions. To address these issues, we created HitPredict (http://hintdb.hgc.jp/htp/), a database of quality assessed interactions in nine species. HitPredict combines interactions from IntAct, BioGRID and HPRD and determines the confidence level of the interactions based on a reliability score calculated using the sequence, structure and functional annotations of the interacting proteins. HitPredict was first introduced in 2005 as a database of high confidence PPIs from high-throughput data sets. It has since been updated annually and has now been expanded to include small-scale interactions along with a more intuitive user interface. Similarly,

Hintdb, a database of homologous protein-protein interactions was also updated.

8. Functional characterization of intrinsically disordered regions in mammalian proteins

Ashwini Patil, Shunsuke Teraguchi[4], Daron Strandley[4] and Kenta Nakai

Intrinsically disordered regions in proteins are regions without a stable tertiary structure. Despite the lack of a stable structure, these regions play an important role in protein function as a result of their flexibility and adaptability. Intrinsically disordered regions are found in proteins functioning in cell signaling and transcription regulation. However, several such regions are not associated with any function and are often ignored during the functional annotation of proteins. In this project, we group similar intrinsically disordered regions in an attempt to identify those that function in a similar manner. We find that disordered regions can be classified into different groups based on the prevalence of charged and hydrophobic residues. We also find that disordered regions in distinct groups are associated with distinct functions suggesting a possible functional classification of disordered regions.

9. Using gene expression correlations to separate functionally distinct genes

Ashwini Patil, Kenta Nakai and Kengo Kinoshita[5]: [5]Tohoku University

The expression patterns of genes under various conditions have been extensively studied. However, the differences in the levels of expressions of genes with different functions have not yet been extensively studied. In this project, we tried to identify if functionally distinct genes had different expression patterns using human gene expression data and a gene expression stability measure. The stability measure indicates how variable the expression of a gene is across multiple samples. We grouped the genes into 4 groups based on their stability and average correlation coefficient and studied the functional differences between them.

Publications

Sathira, N., Yamashita, R., Tanimoto, K., Kanai, A., Arauchi, T., Kanematsu, S., Nakai, K., Suzuki, Y. and Sugano, S. Characterization of transcription start sites of putative non-coding RNAs by multifaceted use of massively paralleled sequencer. DNA Res. 17(3): 169-183, 2010.

Tanaka, Y., Yoshimura, I. and Nakai, K. Positional variations among heterogeneous nucleosome maps give dynamical information on chromatin. Chromosoma. 119(4): 391-404, 2010.

Kubo, A., Suzuki, N., Yuan, X., Nakai, K., Satoh, N., Imai, K. and Satou, Y. Genomic cis-regulatory networks in the *Ciona intestinalis* embryo. Development. 137(10): 1613-1623, 2010.

Patil, A., Kinoshita, K. and Nakamura, H. Hub promiscuity in protein-protein interaction networks. Int. J. Mol. Sci. 11(4): 1930-1943, 2010.

Tanaka, Y., Yamashita, R., Suzuki, Y. and Nakai, K. Effects of Alu elements on global nucleosome positioning in the human genome. BMC Genomics. 11: 309, 2010.

Kawaki, H., Kubota, S., Aoyama, E., Fujita, N., Hanagata, H., Miyauchi, A., Nakai, K. and Takigawa, M. Design and utility of CCN2 anchor peptide aptamers. Biochemie. 92(8): 1010-1015, 2010.

Patil, A., Kinoshita, K. and Nakamura, H. Domain distribution and intrinsic disorder in hubs in the human protein-protein interaction network. Protein Sci. 19(8): 1461-1468, 2010.

Schönbach, C., Nakai, K., Tan, T.-W. and Ranganathan, S. InCoB2010-9th International Conference on Bioinformatics at Tokyo, Japan, September 26-28, 2010. BMC Bioinformatics. 11(Suppl 7): S1, 2010.

Okamura, K., Matsumoto, K. and Nakai, K. Gradual transition from mosaic to global DNA methylation patterns during deuterostome evolution. BMC Bioinformatics. 11(Suppl 7): S2, 2010.

Teraguchi, S.*, Patil, A.* and Standley, D.M. Intrinsically disordered domains deviate significantly from random sequences in innate immune and generic mammalian proteins. BMC Bioinformatics. 11(Suppl 7): S7, 2010. (*joint first author).

Satoh, T., Takeuchi, O., Vandenbon, A., Yasuda, K., Tanaka, Y., Kumagai, Y., Miyake, T., Matsushita, K,. Okazaki, T., Saitoh, T., Honma, K., Matsuyama, T., Yui, K., Tsujimura, T., Standley, D. M., Nakanishi, K., Nakai, K. and Akira, S. The JMJD3-IRF4 axis regulates M2 macrophage polarization and host responses against helminth infection. Nature Immunol. 11(10): 936-944, 2010.

Imai, K. and Nakai, K. Prediction of subcellular location of proteins: where to proceed?. Pro-

teomics. 10(22): 3970-3983, 2010.
Ranganathan, S., Schönbach, C., Nakai, K. and Tan, T.-W. Challenges of the next decade for the Asia Pacific region: 2010 International Conference in Bioinformatics (InCoB 2010). BMC Genomics. 11(Suppl 4): S1, 2010.
Patil, A., Nakai, K. and Nakamura, H. HitPredict: a database of quality assessed protein-protein interactions in nine species. Nucl. Acids Res. 39: D744-D799, 2011. (published online on October 14, 2010.)
Park, S.-J. and Nakai, K. A regression analysis of gene expression in ES cells reveals two gene classes that are significantly different in epigenetic patterns. BMC Bioinformatics 12 (Suppl 1): S50, 2011.
Khare, P., Mortimer, S.I., Cleto, C.L., Okamura, K., Suzuki, Y., Kusakabe, T., Nakai, K., Meedel, T.H. and Hastings, K.E.M. Cross-validated methods for promoter/transcription start site mapping in SL trans-spliced genes, established using the *Ciona intestinalis* troponin I gene. Nucleic Acids Res. published online on November 24, 2010.
中井謙太．ゲノム配列情報解析の課題と未来について．Science Portal China（中国科学技術月報）．2010.

Human Genome Center

Department of Public Policy
公共政策研究分野

Associate Professor	Kaori Muto, Ph.D.	准教授	博士(保健学)	武 藤 香 織
Assistant Professor	Yusuke Inoue, Ph.D.	助 教	博士(社会健康医学)	井 上 悠 輔
Project Assistant Professor	Hyunsoo Hong, Ph.D.	特任助教	学術博士	洪 賢 秀
Project Assistant Professor	Ayako Kamisato, M.A.	特任助教	法学修士	神 里 彩 子

The Department of Public Policy works to achieve three major missions: public policy studies of translational research, its application, and its impact on society; research ethics consultation for scientists to comply with ethical guidelines and to build public trust; and development of "minority-centered" scientific communication. By conducting qualitative and quantitative social science study and policy analysis, we facilitate discussion of challenges arising from advances in medical sciences. Furthermore, we study specific ethics issues related to construction of a human biological substances collection, and related to vaccination policy. We also held a Sci/Art Exhibition titled as "Office Bacteria-the Space" at the Medical Science Museum as one outreach activity undertaken via art.

1. Biobank Japan Project (BBJP) and its ethical, legal and social implications

The Biobank Japan Project (BBJP) is a disease-focused biobanking project headed by Professor Yusuke Nakamura since 2003. Biobank Japan consists of donated DNA, sera and clinical information from 200,000 patients of 66 hospitals in Japan. Informed consent, which ensures the autonomous decisions of participants, is believed to be practically impossible for the biobanking project in general. Consequently, the concept of 'trust', which is "judgment and action in conditions of less than perfect information", has been suggested to compensate for this limitation. To clarify the role and significance of trust, we conducted a questionnaire survey of research coordinators ($n=157$) and of participants ($n=1,378$) in 2007. We also conducted interview studies of research coordinators ($n=50$) in 2010.

Our latest paper (#1) emphasizes the importance of communication with participants after receiving their consent in the case of the biobanking project. After describing the limitations of informed consent within the BBJP based on a survey we conducted, we introduced our attempts to communicate with participants, discussing their implications as a means to compensate for the limitations of informed consent in the biobanking project.

As a means to maintain the participants' trust of the project, research coordinators who had been specially trained for the BBJP have played important roles. We have worked steadily to complete analyses of the research coordinators of the BBJP. At the beginning of the BBJP, their primary roles were recruitment. After the end of recruitment, their roles shifted to the tracing of participants to extract clinical information and to input it into the database. However, their support and encouragement of participants complemented the contents of the initial consent process and reinforced participants' incentive to continue in their role. The results of this study will contribute to improved quality control and

better communication between administrators of long-term research projects and the project participants.

2. Rethinking academic freedom

How should we regulate scientific research, especially life science research, which is developing rapidly? To find a clue to answer this question, Kamisato's paper (#6) returns to article 23 of the Constitution: "Academic Freedom". She described the necessity of reinterpreting this article in response to the demands of the times because it traditionally has not been provided and discussed that Academic Freedom includes the freedom of scientific research. Then, she discussed whether freedom of scientific research is protected by this article, and considered the best ways to regulate scientific research, along with new issues that emerged from the freedom to publish the research results.

3. Research ethics consultation

Public interest in research ethics has grown. Society increasingly makes demands in this area. Provision of a system that can support "on-site" researchers to avail themselves of immediate consultation when concerns and issues arise related to research ethics and other matters is also among those demands. Based on these demands, in recent years, the universities and research institutes providing "research ethics consultation" have become increasingly numerous in the United States. Several papers from the USA explaining "research ethics consultation" reveal that the contents and methods of "research ethics consultation" differ among institutions, and also show that pre-screening of ethical review applications is excluded from the contents of "research ethics consultation" in the US. Our paper (#7) presented analyses of all consulted cases ($n=234$) from April to December 2009 to the Office of Research Ethics (ORE) of the IMSUT from researchers. We extracted 20 categories of their consultation needs and showed that 12 were related to "how to survive ethical review" and pre-screening of documents for ethical review. What is "research ethics consultation" and what creates these differences? We studied the answers to these questions and found that the extent of academic and social understanding of research ethics' importance is critical information for considering these questions.

4. Science and Art

We have been organizing a series of artwork exhibitions related to scientific knowledge, especially that related to biomedical science, at the Museum of Modern Medical Sciences twice a year since 2009. The "IMSUT Science Art Exhibitions" are intended to provide audiences the opportunity to encounter the field of science from the perspectives of artists, and to present the possibility of subjective reflection on what is considered to be the most objective form of knowledge, through the artworks themselves and through talks given by the artists during the exhibition.

This year, we have organized two exhibitions: "Boundary Face ‹-› 界面空間" (24 February 2010-18 March 2011) and "宇宙(Universe)"(6 January 2011-30 January 2011). The former exhibition of installed works by Hideo Iwasaki and Emiko Inoue, developed in the laboratory of Iwasaki, who is also a molecular biologist, received approximately 500 visitors. The latter exhibition of OFFICE BACTERIA is still on view, including oil paintings by Teppei Ikehira and accessories designed by Kanae Briandet, inspired by microscopic pictures taken by Roman Briandet, who is a molecular biologist. Additionally, we introduced our activities as an effective form of science communication by reflecting on the first exhibition we organized in 2008, at the Annual Meeting of the Society for Social Studies of Science (August 25-29, 2010).

As the principle for promotion of dialogues with the public on science and technology, as recently published by the minister in charge of the governance of science and technology, explaining the implications of scientific studies to the public is now regarded as an important responsibility of scientists. Various means of communicating science have been introduced as ways to fulfill this responsibility. With our activities of exhibiting science art, we are proposing another mode of communicating science that covers the limitations of the orthodox mode of science communication based on the dialogue between experts and non-experts.

5. Living liver organ donation in Japan

The global criticism of organ trafficking and transplant tourism requires many countries to pursue legal protection of living organ donors for organ transplantation. Japan is one such criticized country: more than 26,000 people have become living organ donors. Muto's paper (#4) presents an exploration of living liver transplantation in Japan from legal, social and ethical perspectives. Since the first living liver transplantation in 1989, the cases have increased, with extremely high dependency in spite of a few deaths and cases of severe disability. Govern-

ment and professional guidelines stipulate that living donors be "relatives" so that living organ transplantation can be privatized and regarded as a family issue, although it is strictly limited to altruistic cases in some countries. Based on results of the Living Liver Donor Survey conducted in 2004, Japanese liver donors have had varied experiences. Most male donors were employed, felt some obligation, and harbored concerns about financial effects and employment during decision-making. In contrast, only a quarter of female donors were employed, felt guilty about the health conditions of their children, and did not have opportunities for regular health checkups after donation. Severe tensions and family dissolution occurred in adult-to-adult cases, although donors were satisfied with donation overall. The author suggests that we should rethink privatization of living organ donation and that independent advocates should support potential donors. Further research is necessary to explore the reasons why organ donation is privatized even in some forms of cadaver cases in Japan.

6. Ethical issues in human tissue and data banking

Banking human tissue samples and data for future use (biobanking) is a fundamental component of the infrastructure supporting biomedical research activities. For research activities, collecting these resources has been important for medical research activities, but research ethics in biobanking has some unique characteristics that differ from traditional notions of human subjects' protection in clinical research. First, we have studied consent and monitoring processes related to biobanking activities to explore the ideal relation between society and biobanking. As some results, we presented some identified ethical issues related to the large scale clinical database in summer (Japanese College of Cardiology, September 2010). Second, we investigated ethical issues in cadaveric research. Primarily two theories have prevailed: one regards a body (or a part of it) as a kind of object which can be a property, and the other maintains that a dead body should be respected as a remnant of an individual's personality or identity. Additionally, some explanations suggest that a dead body itself deserves respect. The role of the bereaved family and moral status of the dead body varies depending on these different standpoints. We have just started questionnaire surveys of human tissue collections in forensic medicine laboratories in Japan to elucidate the status of a human body, and to evaluate socially expected requirements for research usage.

7. Vaccination Policy

Striking a balance between the rapid availability of a novel vaccine while ensuring its safety, quality, and efficacy is a major challenge during a pandemic. We elucidated physicians' attitudes related to novel vaccines during the influenza A /H1N1 pandemic of 2009, and to determine factors that affected their vaccination recommendations to patients (#2). Of a random sample of 1,000 general practitioners (GPs) in Japan, 515 participated in a cross-sectional anonymous survey conducted immediately before the novel vaccine became available (between 28 September and 18 October 2009). In all, 453 GPs (88.3%) replied that they intended to receive the new vaccine themselves. However, only 177 GPs (34.6%) intended to recommend it proactively to their patients. The anticipated cost of the vaccine negatively influenced the intention to vaccinate themselves and their recommendations to patients ($P < 0.001$, χ^2test). Results of multivariate logistic regression analysis showed that physicians with experience in influenza A/H1N1 patient contacts [1-20 contacts, odds ratio (OR)= 7.49 (95% confidence interval [CI]: 1.73-32.36), $P = 0.007$; >20 contacts, OR=8.03 (95% CI: 1.77-36.50), $P = 0.007$, compared with no contacts] were more likely to recommend the vaccine to patients, although those with knowledge of the fear of a causal association between Guillain-Barre syndrome (GBS) cases and the 1976 swine flu vaccination in the USA were less likely to recommend the vaccine [OR=0.66 (95% CI: 0.45-0.97), $P = 0.036$]. Results of our survey indicate that physicians experience a moral conflict related to the recommendation of the novel vaccine to patients, which might result from their own experience with the disease, knowledge of vaccine side-effects, and cost.

8. Research in progress

We have been conducting other studies as described below.

- Ethical, legal and social implications of commercial genetic/genomic testing services in eastern Asia
- Development and evaluation of communication methods with participants of Biobank Japan and other long-term studies
- Analysis of roles of research coordinators for better recruitment and for building trust from participants
- Ethical, legal and social implications of stem cell studies including animal-human chimeric embryos and iPS cell banking
- Bench-side research ethics consultation and quality assurance of research ethics committees

- Science communication through art and ethical challenges of biomaterial art
- Ethics issues in collecting human biological substances for constructing a research infrastructure
- Vaccination policy

Publications

1. Watanabe M, Inoue Y, Chang C, Hong H, Kobayashi I, Suzuki S, Muto K. For what am I participating?-The need for communication after receiving consent form biobanking project participants: experience in Japan. J. Human Genetics, in press, 2011.
2. Inoue Y, Matsui K. Physicians' recommendations to their patients concerning a novel vaccine: a cross-sectional survey on 2009 A/H1N1 vaccination in Japan. Environmental Health and Preventive Medicine, in press, 2011.
3. Inoue Y, Wada Y, Motohashi Y, Koizumi A. History of blood transfusion before 1990 increases cancer mortality risk independent of liver diseases: prospective long-term follow-up. Environmental Health and Preventive Medicine, 15: 180-187, 2010.
4. Muto K. Organ transplantation as a family issue: living liver donors in Japan. International Journal of Japanese Sociology, 19(1): 35-48, 2010.
5. Semba Y, Chang C, Hong H, Kamisato A, Kokado M, Muto K. Surrogacy: donor conception regulation in Japan. Bioethics, 24(7): 348-57, 2010.
6. 神里彩子．科学研究規制をめぐる「学問の自由」の現代的意義と課題．社会技術研究論文集，7，211–221，2010.
7. 神里彩子，武藤香織．「研究倫理コンサルテーション」の現状と今後の課題—東京大学医科学研究所研究倫理支援室の経験より．生命倫理：21，183–193，2010.
8. 井上悠輔，赤林朗．発生学の展開と幹細胞研究の諸問題．Biophilia：6，66–70，2010.
9. 井上悠輔．人体要素を研究資源として利用する際の研究倫理上の諸問題．医療・生命と倫理・社会：9，1–22，2011.
10. 玉腰暁子，武藤香織．『医療現場における調査研究倫理ハンドブック』，東京：医学書院，2011.