*Human Genome Center*

# Laboratory of Genome Database
# Laboratory of Sequence Analysis
# ゲノムデータベース分野
# シークエンスデータ情報処理分野


| | | | |
|---|---|---|---|
| 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 |
| Lecturer | Tetsuo Shibuya, Ph.D. | 講師 理学博士 | 渋谷哲朗 |
| Assistant Professor | Michihiro Araki, Ph.D. | 助教 薬学博士 | 荒木通啓 |



*DNA, RNA, and proteins are the basic molecular building blocks of life, but the living cell contains additional molecules, including water, ions, small chemical compounds, glycans, lipids, and other biochemical molecules, without which the cell would not function. We are developing bioinformatics methods to integrate different types of data and knowledge on various aspects of the biological systems towards basic understanding of life as a molecular interaction/reaction system and also for practical applications in medical and pharmaceutical sciences.*


## 1. KEGG DRUG and KEGG DISEASE


**Minoru Kanehisa and Michihiro Araki**


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 or the organism from genomic information. Although the basic components of the KEGG resource are developed in Kyoto University, this Laboratory in the Human Genome Center is responsible for the applied areas of KEGG, especially in medical and pharmaceutical sciences. We develop KEGG DRUG (http://www.genome.jp/kegg/drug/), which is a chemical structure database for all approved drugs, associated with target information in the context of KEGG pathways, efficacy information in the context of hierarchical drug classifications, etc. We also develop KEGG DISEASE (http://www.genome.jp/kegg/disease/) as a new addition to the KEGG suite of databases. Each disease entry consists of a list of diseases genes and other lists of molecules such as environmental factors, markers, drugs, etc. Both DRUG and DISEASE are highly integrated with other KEGG databases including PATHWAY, BRITE, GENES, and COMPOUND, and also with other Internet resources.

## 2. Automatic assignments of orthologs and paralogs in complete genomes


**Toshiaki Katayama, Shuichi Kawashima, Akihiro Nakaya and Minoru Kanehisa**


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 2.5 mil-

lion genes from over 600 organisms as of October 2007. Sequence similarities among these genes are calculated by all-against-all SSEARCH comparison and stored them 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 developed a fully automated procedure to find candidate orthologous clusters including whose current functional annotation is anonymous. The method is based on a graph analysis of the SSDB database, treating genes as nodes and Smith-Waterman sequence similarity scores as weight of edges. 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. A web interface to search and browse genes in clusters is made available at http://dev.kegg.jp/oc/.

## 3. EGENES: Transcriptome-Based Plant Database of Genes with Metabolic Pathway Information and Expressed Sequence Tag Indices in KEGG

**Ali Masoudi-Nejad, Susumu Goto, Ruy Jauregui, Masumi Itoh, Shuichi Kawashima, Yuki Moriya, Takashi R. Endo and Minoru Kanehisa**

EGENES is a knowledge-based database for efficient analysis of plant expressed sequence tags (ESTs) that was recently added to the KEGG suite of databases. It links plant genomic information with higher order functional information in a single database. It also provides gene indices for each genome. The genomic information in EGENES is a collection of EST contigs constructed from assembled ESTs by using EGassembler. EGassembler is a web server, which provides an automated as well as a user-customized analysis tool for cleaning, repeat masking, vector trimming, organelle masking, clustering and assembling of ESTs and genomic fragments. Due to the extremely large genomes of plant species, the bulk collection of data such as ESTs is a quick way to capture a complete repertoire of genes expressed in an organism. EGENES and EGassembler are publicly available at http://www.genome.jp/kegg-bin/create_kegg_menu?category5plants_egenes and http://egassembler.hgc.jp/respectively.

## 4. SOAP/WSDL interface for the KEGG system

**Shuichi Kawashima, Toshiaki Katayama and Minoru Kanehisa**

KEGG (Kyoto Encyclopedia of Genes and Genomes) 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), the information about the genomic space (GENES database), and information about the chemical space (LIGAND and DRUG databases). To facilitate large-scale applications of the KEGG system programatically, we have been developing and maintaining the KEGG API as a SOAP/WSDL based web service. Recent improvements includes retrieval of the reference information from the KEGG PATHWAY database and utilization of the newly developed KEGG DRUG database. We are refining our service to compromise standardization of the bioinformatics web services and checking of operations in various computer languages. The KEGG API is available at http://www.genome.jp/kegg/soap/.

## 5. Comprehensive repository for community genome annotation

**Toshiaki Katayama, Mari Watanabe and Minoru Kanehisa**

KEGG DAS is an advanced genome database system providing DAS (Distributed Annotation System) service for all organisms in the GENOME and GENES databases in KEGG (Kyoto Encyclopedia of Genes and Genomes). Currently, KEGG DAS contains over 8 million annotations assigned to the genome sequences of 615 organisms (increased from 440 organisms in last year). The KEGG DAS server provides gene annotations linked to the KEGG PATHWAY and LIGAND databases. In addition to the coding genes, information of non-coding RNAs predicted using Rfam database is also provided to fill the annotation of the intergenic regions of the genomes. We have been developing the server based on open source software including BioRuby, BioPerl, BioDAS and GMOD/GBrowse to make the system consistent with the existing open standards. We are also responsible to the Japanese localization of the GBrowse genome browser. The contents of the KEGG DAS data-

base can be accessed graphically in a web browser using GBrowse GUI (graphical user interface) and also programatically by the DAS (XML over HTTP) protocol. Users are able to add their own annotations onto the KEGG DAS server by connecting other DAS servers or by simply uploading their own data as a file. This functionality enables researchers to share the "community annotation." The KEGG DAS is weekly updated and freely available at http://das.hgc.jp/.

## 6. Constructing a database of full length cDNA of pathogenetic arthropods

**Toshiaki Katayama, Shuichi Kawashima, Junichi Watanabe, Yutaka Suzuki, Sumio Sugano and Minoru Kanehisa**

Anopheles mosquito, tsetse fly, tsutsugamushimite, dust mite are arthropods which are known as medically important because these either transmit various infectious disease including malaria, Japanese river fever, or cause allergy such as asthma and dermatitis. Because of serious medical problems they cause, their genomes are being extensively analyzed recently. We have produced libraries of the four organisms and are constructing their databases for the functional genome analysis. It will be available on the site http://fullarth.hgc.jp/

## 7. AAindex: Amino Acid index database

**Shuichi Kawashima, Toshiaki Katayama and Minoru Kanehisa**

AAindex is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids. We have added a collection of protein contact potentials to the AAindex as a new section. Accordingly AAindex consists of three sections now: AAindex1 for the amino acid index of 20 numerical values, AAindex2 for the amino acid mutation matrix and AAindex3 for the statistical protein contact potentials. All data are derived from published literature. The database can be accessed through the DBGET/LinkDB system at GenomeNet (http://www.genome.jp/dbget-bin/www_bfind?aaindex) or downloaded by anonymous FTP (ftp://ftp.genome.jp/pub/db/community/aaindex/).

## 8. Comparative pair-wise domain-combination for screening the clade specific domain-architectures in metazoan genomes

**Shuichi Kawashima, Takeshi Kawashima, Hiroshi Wada and Minoru Kanehisa**

In the evolution of the eukaryotic genome, exon or domain shuffling has produced a variety of proteins. On the assumption that each fusion event between two independent protein-domains occurred only once in the evolution of metazoans, we can roughly estimate when the fusion events were happened. For this purpose, we made phylogenetic profiles of pair-wise domain-combinations of metazoans. The phylogenetic profiles can be expected to reflect the protein evolution of metazoan. Interestingly, the phylogenetic tree of metazoans, derived from the profiles, supported the "Ecdysozoa hypothesis" that is one of the major hypotheses for metazoan evolution. Further, the phylogenetic profiles showed the candidates of genes that were required for each clade-specific features in metazoan evolution. We propose that comparative proteome analysis focusing on pair-wise domain-combinations is a useful strategy for researching the metazoan evolution. Additionally, we found that the extant ecdysozoans share only fourteen domain-combinations in our profiles. Such a small number of ecdysozoan-specific domain-combinations is consistent with the extensive gene losses through the evolution of ecdysozoans.

## 9. SSS: a sequence similarity search service

**Toshiaki Katayama, Kazuhiro Ohi and Minoru Kanehisa**

There are various services in the world to find similar sequences from the database, such as the famous BLAST service provided at NCBI. However, the method to search and the database to be searched could not be added from outside. To provide our super computer resources at the Human Genome Center to the research community, we started to develop a new service for the sequence similarity search, SSS. In SSS, user can select the search algorithm from BLAST, FASTA, SSEARCH, TRANS and EXONERATE. This variety of options is unique among the public services. Then user is prompted to select appropriate database depending on the algorithm selected and the search is executed. On the backend, we implemented the search system on the Sun Grid Engine to provide efficient resources on distributed computers. As a result, we are able to provide time consuming services such as TRANS and EXONERATE in addition to the popular algorithms. The SSS service is freely available at http://sss.hgc.jp/.

## 10. High performance database entry retrieval system

**Toshiaki Katayama, Shuichi Kawashima, Kenta Nakai and Minoru Kanehisa**

Recently, 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 83,051,496 (Release 162+ 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. The HiGet system is constructed on the HiRDB, a commercial ORDBMS (Object-oriented Relational Database Management System) developed by Hitachi, Ltd. It is publicly accessible on the Web page at http://higet.hgc.jp/ or SOAP based web service at http://higet.hgc.jp/soap/. HiGet can execute full text search on various biological databases. In addition to the original plain format, the system contains data in the XML format in order to provide a field specific search facility. When a complicated search condition is issued to the system, the search processing is executed efficiently by combining several types of indices to reduce the number of records to be processed within the system. Current searchable databases are GenBank, UniProt, Prosite, OMIM, PDB and RefSeq. We are planning to include other valuable databases and also planning to develop an inter-database search interface and a complex search facility combining keyword search and sequence similarity search.

## 11. PSGST: A New Data Structure for Indexing Protein 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. The geometric suffix tree, which is also proposed by our group, is a very sophisticated index structure that enables fast and accurate search on protein 3-D structures. By using it, we can search from 3-D structure databases for all the substructures whose RMSDs (root mean square deviations) to a given query 3-D structure are not larger than a given bound. We proposed a new data structure based on the geometric suffix tree whose query performance is much better than the original geometric suffix tree. We call the modified data structure the prefix-shuffled geometric suffix tree (or PSGST for short). According to our experiments, the PSGST outperforms the geometric suffix tree in most cases. The PSGST shows its best performance when the database does not have many substructures similar to the query. The query is sometimes 100 times faster than the original geometric suffix trees in such cases.

## 12. Compact Geometric Suffix Tree

**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. The geometric suffix tree, which is also a data structure proposed by our group, is a very sophisticated index structure that enables fast and accurate search on protein 3-D structures. By using it, we can search from 3-D structure databases for all the substructures whose RMSDs (root mean square deviations) to a given query 3-D structure are not larger than a given bound. But the geometric suffix tree requires rather large memory, i.e., about 65n byte for a protein of length n. We developed a new data structure called "compact geometric suffix tree" which requires only 23n byte, which is almost about 1/3 of the original geometric suffix tree, while the query speed is almost the same as the original geometric suffix tree.

## 13. Protein Function Prediction based on 3-D Structure Motifs

**Chia-Han Chu, Hiroki Sakai, and Tetsuo Shibuya**

Protein functions are said to be determined by its 3-D structures, but not all functions have been known to be related to some 3-D structure motifs. The geometric suffix tree, a data structure for indexing 3-D protein structures, which is also developed by us, enables comprehensively enumeration of all the possible structural motifs among given set of proteins. We are developing a new algorithm based on the support vector machine that decides protein's function from the 3-D structure of a protein. This algorithm utilizes all the possible 3-D motifs found by using the geometric suffix tree.

## 14. Fast Hinge Detection Algorithm in Protein Structures

**Tetsuo Shibuya**

Analysis of conformational changes is one of the keys to the understanding of protein functions and interactions. For the analysis, we often compare two protein structures, taking flexible regions like hinge domains into consideration. The RMSD (Root Mean Square Deviation) is the most popular measure for comparing two protein structures, but it is only for rigid structures without hinge domains. In this paper, we propose a new measure called RMSDh (Root Mean Square Deviation considering hinges) and its variant RMSDh(k) for comparing two flexible proteins with hinge domains. We also propose novel efficient algorithms for computing them, which can detect the hinge positions at the same time. The RMSDh is suitable for cases where there is one small hinge domain in each of the two target structures. The new algorithm for computing the RMSDh runs in linear time, which is same as the time complexity for computing the RMSD and is faster than any of previous algorithms for hinge detection. The RMSDh(k) is designed for comparing structures with more than one hinge domain. The RMSDh (k) measure considers at most k small hinge domains, i.e., the RMSDh(k) value should be small if the two structures are similar except for at most k hinge domains. To compute the value, we propose an $O(kn^2)$-time and O(n)-space algorithm based on a new dynamic programming technique. We also test our measures against both flexible protein structures and non-flexible protein structures, and show that the hinge positions can be correctly detected by our algorithms.

## 15. Fast Flexible Protein Structure Alignment

**Kohichi Suematsu and Tetsuo Shibuya**

The Hinge Detection Algorithm described in section 14 only considered rigid hinge points, but the hinges are sometimes bends a little by itself, which sometimes leads to inaccurate prediction of hinge positions. Thus we incorporated the notion 'bending hinge' to detect such hinge positions. We developed a very efficient heuristic algorithm for finding such bending hinges, as the exact algorithm for this problem requires exponential time.

## 16. Fast Algorithms for Fast Range RMSD Query

**Tetsuo Shibuya**

Protein structure analysis is a very important research topic in the molecular biology of the post-genomic era. The RMSD (root mean square deviation) is the most frequently used measure for comparing two protein 3-D structures. We deal with two fundamental problems related to the RMSD. We first deal with a problem called the 'range RMSD query' problem. Given an aligned pair of structures, the problem is to compute the RMSD between two aligned substructures of them without gaps. This problem has many applications in protein structure analysis. We propose a linear-time preprocessing algorithm that enables constant-time RMSD computation. Next, we consider a problem called the 'substructure RMSD query' problem, which is a generalization of the above range RMSD query problem. It is a problem to compute the RMSD between any substructures of two unaligned structures without gaps. Based on the algorithm for the range RMSD problem, we propose an O(nm) preprocessing algorithm that enables constant-time RMSD computation, where n and m are the lengths of the given structures. Moreover, we propose O(nm log r/r) -time and O(nm/r)-space preprocessing algorithm that enables O(r) query, where r is an arbitrary integer such that $0 < r < \min(n, m)$. We also show that our strategy also works for another measure called the URMSD (unit-vector root mean square deviation), which is a variant of the RMSD.

## 17. Suffix Array Construction with a Lazy Scheme

**Ben Hachimori and Tetsuo Shibuya**

The suffix array is one of the most important indexing data structures for alphabet strings, including DNA sequences, RNA sequences, protein sequences, web pages, Medline database, and so on. But even the most sophisticated algorithm for constructing the suffix array requires a lot of time. We developed a new efficient lazy algorithm that computes the suffix array only after we get the query. By doing so, we have to compute only the necessary part of the suffix array.

## 18. Genotype Clustering based on Hidden Markov Models

**Ritsuko Onuki, Tetsuo Shibuya and Minoru Kanehisa**

Haplotype clustering is important for gene mapping of human disease. Although its importance for the analysis, it is difficult to obtain haplotype data from present experiment for its cost and error rate. Instead of haplotypes, genotypes are much easier to obtain. In this work,

we propose a new method for clustering genotypes. In the algorithm. we first infer the multiple haplotype candidates from the genotype, and next we calculate the distance between the genotypes based on the results of the haplotype inference. Then we perform genotype clustering based on the distances. We evaluated our algorithm by applying our algorithm against several actual genotype data.

### 19. Analysis of the Chemical Diversity of Medicinal Natural Products

**Michihiro Araki, Tetsuo Shibuya, Kohichi Suematsu, and Minoru Kanehisa**

The diverse activities of medicinal natural products are explained by extensive diversities in the chemical structures. Although several hypotheses have been proposed to understand the origin of the chemical diversity, no systematic analyses have been done so far. In order to elucidate how the chemical diversities are designed in the biosynthetic circuits, we defined molecular building blocks required for describing the chemical information of natural products and performed the systematic analyses with the transformational and combinatorial information of the building blocks. We further reconstructed the metabolic network using the chemical and the genomic information to identify alternative biosynthetic pathways as well as new chemical entities.

### 20. Extraction of Chemical Modification Patterns from Drug Developmental History and Its Application

**Daichi Shigemizu, Michihiro Araki, Shujiro Okuda, Susumu Goto and Minoru Kanehisa**

The chemical modifications in the history of drug discovery are expected to provide a wealth of the medicinal chemists' knowledge and it is of importance to extract the information on the chemical modifications in the form of distinct patterns. KEGG DRUG structure map collected such information by illustrating the developmental relationships among the drug structures. Here we focused on 288 drug pairs in the KEGG DRUG structure map to extract the chemical modification patterns. We developed a method for taking the transformations of core chemotypes and modified fragments into considerations. A set of the pairs of the atoms on the core chemotypes and the fragments was collected as the indicators of the chemical modifications, which identified 347 chemical modification patterns in the structure maps. The transformation patterns were then applied for in silico drug modifications. We developed a scheme to predict potential drug-like compounds, which could demonstrate the reconstruction of the drug developmental maps.

### 21. Comprehensive Analysis of Distinctive Polyketide and Nonribosomal Peptide Structural Motifs Encoded in Microbial Genomes

**Yohsuke Minowa, Michihiro Araki and Minoru Kanehisa**

We developed a highly accurate method to predict polyketide (PK) and nonribosomal peptide (NRP) structures encoded in the microbial genome. PKs/NRPs are polymers of carbonyl or peptidyl chains synthesized by polyketide synthases (PKS) and nonribosomal peptide synthetases (NRPS). We analyzed domain sequences corresponding to specific substrates and physical interactions between PKSs/NRPSs in order to predict which substrates are selected and assembled into highly ordered chemical structures. The predicted PKs/NRPs were represented as the sequences of carbonyl/peptidyl units to extract the structural motifs efficiently. The structural sequences were compared using the Smith-Waterman algorithm to extract 33 structural motifs that are significantly related with their bioactivities. The integrative analysis of genomic and chemical information here will provide a strategy to predict the chemical structures, the biosynthetic pathways, and the biological activities of PKs/NRPs, which is useful for the rational design of novel PKs/NRPs.

### 22. Prediction of Drug-target Interaction Networks from the Integration of Chemical and Genomic Spaces

**Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda and Minoru Kanehisa**

The identification of interactions between drugs and target proteins is a key area in drug discovery. We characterize four classes of drug-target interaction networks in humans involving enzymes, ion channels, GPCRs, and nuclear receptors, and reveal significant correlations between drug structure similarity, target sequence similarity, and the drug-target interaction network topology. We then develop new statistical methods to predict unknown drug-target interaction networks from chemical structure and genomic sequence information simultaneously on a large scale. As a result, we demonstrate the

usefulness of our proposed method for the prediction of the four classes of drug-target interaction networks. Our comprehensively predicted drug-target interaction networks enable us to suggest many potential drug-target interactions and to increase research productivity toward genomic drug discovery.

## Publications


Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. KEGG for linking genomes to life and the environment. *Nucleic Acids Research* 36: 2008, in press.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research* 36: 2008, in press.

Lapp, H., Bala, S., Balhoff, J.P., Bouck, A., Goto, N., Holder, M., Holland, R., Holloway, A., Katayama, T., Lewis, P.O., Mackey, A., Osborne, B.I., Piel, W.H., Kosakovsky Pond, S.L., Poon, A., Qiu, W.-G, Stajich, J.E., Stoltzfus, A., Thierer, T., Vilella, A.J., Vos, R.A., Zmasek, C. M., Zwickl, D. and Vision, T.J. The 2006 NESCent Phyloinformatics Hackathon: A field report, *Evolutionary Bioinformatics*, 2008, in press.

Kawashima, S., Kawashima, T., Putnam, N.H., Rokhsar D.S., Wada, H. and Kanehisa, M. Comparative pair-wise domain-combinations for screening the clade specific domain-architectures in metazoan genomes, *Genome Informatics*, 19: 50-60, 2007

Hu, Z., Ng, D.M., Yamada, T., Chen, C., Kawashima, S., Mellor, J., Linghu, B., Kanehisa, M., Stuart, J.M. and DeLisi, C. VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Research*, 35: W625-32, 2007

Masoudi-Nejad, A., Goto, S., Jauregui, R., Ito, M., Kawashima, S., Moriya, Y., Endo, T.R. and Kanehisa, M. EGENES: transcriptome-based plant database of genes with metabolic pathway information and expressed sequence tag indices in KEGG. *Plant Physiology*, 144: 857-866, 2007

Okamoto, S., Yamanishi, Y., Ehira, S., Kawashima, S., Tonomura, K. and Kanehisa, M. Prediction of nitrogen metabolism-related genes in Anabaena by kernel-based network analysis. *Proteomics*, 7: 900-909, 2007

Okuda, S., Kawashima, S., Kobayashi, K., Ogasawara, N., Kanehisa, M. and Goto, S. Characterization of relationships between transcriptional units and operon structures in Bacillus subtilis and Escherichia coli, *BMC Genomics*, 8: 48, 2007

Huang, J., Kawashima, S. and Kanehisa, M. New amino acid indices based on residue network topology, *Genome Informatics*, 18, 2007, in press

Shibuya, T., Combitorial Pattern Matching for Protein 3-D Structures, SIG FPAI, 2008, to appear.

Minowa, Y., Araki, M., Kanehisa, M. Comprehensive Analysis of Distinctive Polyketide and Nonribosomal Peptide Structural Motifs Encoded in Microbial Genomes, Journal of Molecular Biology, 368: 1500-1517, 2007.

Shibuya, T., Efficient Substructure RMSD Query Algorithms, Journal of Computational Biology, Vol. 14, No. 9., 2007, pp. 1201-1207.

Shibuya, T., Prefix-Shuffled Geometric Suffix Tree, Proc. 14th String Processing and Information Retrieval Symposium (SPIRE 2007), LNCS 4726, 2007, pp. 300-309.

Shibuya, T., Current Bioinformatics and Its Future, Journal of the Institute of Electronics, Information and Communication Engineers, vol. 90, no. 2, 2007, pp. 145-147.

Onuki, R., Shibuya, T., and Kanehisa, M., Haplotype Inference Probability-based Genotype Clustering, Proc. 7th Annual Workshop on Bioinformatics and Systems Biology (IBSB), 2007, pp. 26-27.

Shibuya, T., Efficient Substructure RMSD Query Algorithms, IPSJ SIG Notes SIGAL 114-8, 2007, pp. 57-64.

Shibuya, T., Fast and Accurate Algorithms for Protein Hinge Detection, IPSJ SIG Notes SIGBIO 10-4, 2007, pp. 25-32.

Shibuya, T., Prefix-Shuffled Geometric Suffix Tree, IPSJ SIG Notes SIGAL 112-1, May 11, 2007, pp. 1-8.

Onuki, R., Shibuya, T., and Kanehisa, M., Genotype Clustering based on Haplotype Inference, Japanese Society of Human Genetics, 2007.

Onuki, R., Shibuya, T., and Kanehisa, M., A New Method for Genotype Clustering based on Haplotype Inference, Biochemistry and Molecular Biology (BMB 2007), 1P-1071, 2007.

渋谷哲朗，坂内英夫(訳)，Neil C. Jones, Pavel Pevzner(著)，バイオインフォマティクスのためのアルゴリズム入門，共立出版，2007.

*Human Genome Center*

# Laboratory of DNA Information Analysis
## DNA情報解析分野


| | | | | | |
|---|---|---|---|---|---|
| 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　﨑　正　朗 |
| Project Lecturer | Rui Yamaguchi, Ph.D. | | 特任講師 | 博士(理学) | 山　口　　類 |



*The original aim of the research at this laboratory is to establish computational methodologies for discovering and interpreting information of nucleic acid sequences, proteins and some other experimental data arising from researches in Genome Science. 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. Computational Systems Biology

#### a. Cell System Ontology: Representation for modeling, visualizing, and simulating biological pathways


**Euna Jeong, Masao Nagasaki, Ayumu Saito, Satoru Miyano**


With the rapidly accumulating knowledge of biological entities and networks, there is a growing need for a general framework to understand this information at a system level. In order to understand life as system, a formal description of system dynamics with semantic validation will be necessary. Within the context of biological pathways, several formats have been proposed, e.g., SBML, CellML, and BioPAX. Unfortunately, these formats lack the formal definitions of each term or fail to capture the system dynamics behavior. Thus, we have developed a new system dynamics centered ontology called Cell System Ontology (CSO). As an exchange format, the ontology is implemented in the Web Ontology Language (OWL), which enables semantic validation and automatic reasoning to check the consistency of biological pathway models. The features of CSO are as follows: (1) manipulation of different levels of granularity and abstraction of pathways, e.g., metabolic pathways, regulatory pathways, signal transduction pathways, and cell-cell interactions; (2) capture of both quantitative and qualitative aspects of biological function by using hybrid functional Petri net with extension (HFPNe); and (3) encoding of biological pathway data related to visualization and simulation, as well as modeling. The new ontology also predefines mature core vocabulary, which will be necessary for creating models with system dynamics. In addition, each of the core terms has at least one standard icon for easy modeling and accelerating the exchangeability among applications. In order to demonstrate the potential of CSO-based pathway modeling, visualization, and simulation, we present an HFPNe model for the ASEL and

ASER regulatory networks in *Caenorhabditis elegans*.

### b. Conversion from BioPAX to CSO for system dynamics and visualization of biological pathway

**Euna Jeong, Masao Nagasaki, Satoru Miyan**

The vast accumulation of biological pathway data scattered in various sources presents challenges in the exchange and integration of these data. Major new standards for representation of pathway data and the ability to check inconsistency in pathways are inevitable for the development of a reliable pathway data repository. Within the context of biological pathways, the cell system ontology (CSO) had been developed as a general framework to model system dynamics and visualization of diverse biological pathways. CSO provides an excellent environment for modeling, visualizing, and simulating complex molecular mechanisms at different levels of details. This paper examines whether CSO addresses the integration capability of pathway data with system dynamics. We present a conversion tool for converting BioPAX to CSO. Transforming the data from BioPAX to CSO not only allows an analysis of the dynamic behaviors in molecular interactions but also allows the results to be stored for further biological investigations, which is not possible in BioPAX. The conversion is done using simple inference algorithms with the addition of view- and simulation-related properties. We demonstrate how CSO can be used to build a complete and consistent pathway repository and enhance the interoperability among applications.

### c. An efficient grid layout algorithm for biological networks utilizing various biological attributes

**Kaname Kojima, Masao Nagasaki, Euna Jeong, Mitsuru Kato, Satoru Miyano**

Clearly visualized biopathways provide a great help in understanding biological systems. However, manual drawing of large-scale biopathways is time consuming. We proposed a grid layout algorithm that can handle gene-regulatory networks and signal transduction pathways by considering edge-edge crossing, node-edge crossing, distance measure between nodes, and subcellular localization information from Gene Ontology. Consequently, the layout algorithm succeeded in drastically reducing these crossings in the apoptosis model. However, for larger-scale networks, we encountered three problems: (i) the initial layout is often very far from any local optimum because nodes are initially placed at random, (ii) from a biological viewpoint, human layouts still exceed automatic layouts in understanding because except subcellular localization, it does not fully utilize biological information of pathways, and (iii) it employs a local search strategy in which the neighborhood is obtained by moving one node at each step, and automatic layouts suggest that simultaneous movements of multiple nodes are necessary for better layouts, while such extension may face worsening the time complexity. We propose a new grid layout algorithm. To address problem (i), we devised a new forcedirected algorithm whose output is suitable as the initial layout. For (ii), we considered that an appropriate alignment of nodes having the same biological attribute is one of the most important factors of the comprehension, and we defined a new score function that gives an advantage to such configurations. For solving problem (iii), we developed a search strategy that considers swapping nodes as well as moving a node, while keeping the order of the time complexity. Though a naive implementation increases by one order, the time complexity, we solved this difficulty by devising a method that caches differences between scores of a layout and its possible updates. Layouts of the new grid layout algorithm are compared with that of the previous algorithm and human layout in an endothelial cell model, three times as large as the apoptosis model. The total cost of the result from the new grid layout algorithm is similar to that of the human layout. In addition, its convergence time is drastically reduced (40% reduction).

### d. Modelling and simulation of signal transductions in an apoptosis pathway by using timed Petri nets

**Chen Li, Qi-Wei Ge[1], Mitsuru Nakata[1], Hiroshi Matsuno[1], Satoru Miyano: [1]Yamaguchi University**

This paper first presents basic Petri net components representing molecular interactions and mechanisms of signalling pathways, and introduces a method to construct a Petri net model of a signalling pathway with these components. Then a simulation method of determining the delay time of transitions, by using timed Petri nets, i.e. the time taken in firing of each transition is proposed based on some simple principles that the number of tokens flowed into a place is equivalent to the number of tokens flowed out. Finally, the availability of proposed method is confirmed by observing signalling

transductions in biological pathways through simulation experiments of the apoptosis signalling pathways as an example.

### e. Understanding endothelial cell apoptosis: What can the transcriptome glycome and proteome reveal?

**Muna Affara[2], Benjamin Dunmore[2], Christopher Savoie[3], Seiya Imoto, Yoshinori Tamada[3], Hiromitsu Araki[3], D. Stephen Charnock-Jones[2], Satoru Miyano, Cristin Print[4]: [2]Cambridge University, [3]GNI, Ltd. [4]University of Auckland**

Endothelial cell (EC) apoptosis may play an important role in blood vessel development, homeostasis and remodelling. In support of this concept, EC apoptosis has been detected within remodelling vessels in vivo, and inactivation of EC apoptosis regulators has caused dramatic vascular phenotypes. EC apoptosis has also been associated with cardiovascular pathologies. Therefore, understanding the regulation of EC apoptosis, with the goal of intervening in this process, has become a current research focus. The protein-based signalling and cleavage cascades that regulate EC apoptosis are well known. However, the possibility that programmed transcriptome and glycome changes contribute to EC apoptosis has only recently been explored. Traditional bioinformatic techniques have allowed simultaneous study of thousands of molecular signals during the process of EC apoptosis. However, to progress further, we now need to understand the complex cause and effect relationships among these signals. In this article, we will first review current knowledge about the function and regulation of EC apoptosis including the roles of the proteome transcriptome and glycome. Then, we assess the potential for further bioinformatic analysis to advance our understanding of EC apoptosis, including the limitations of current technologies and the potential of emerging technologies such as gene regulatory networks.

### f. Weighted Lasso in graphical Gaussian modeling for large gene network estimation based on microarray data

**Teppei Shimamura, Seiya Imoto, Rui Yamaguchi, Satoru Miyano**

We propose a statistical method based on graphical Gaussian models for estimating large gene networks from DNA microarray data. In estimating large gene networks, the number of genes is larger than the number of samples, we need to consider some restrictions for model building. We propose weighted lasso estimation for the graphical Gaussian models as a model of large gene networks. In the proposed method, the structural learning for gene networks is equivalent to the selection of the regularization parameters included in the weighted lasso estimation. We investigate this problem from a Bayes approach and derive an empirical Bayesian information criterion for choosing them. Unlike Bayesian network approach, our method can find the optimal network structure and does not require to use heuristic structural learning algorithm. We conduct Monte Carlo simulation to show the effectiveness of the proposed method. We also analyze Arabidopsis thaliana microarray data and estimate gene networks.

### g. A structure learning algorithm for inference of gene networks from microarray gene expression data using Bayesian networks

**Kazuyuki Numata, Seiya Imoto, Satoru Miyano**

Estimation of gene networks based on microarray gene expression data is an important problem in systems biology. In this paper we use Bayesian networks as a mathematical model for reverse-engineering gene networks from microarray data. In such a case, structural learning of Bayesian networks is known as an NP-hard problem and we need to use heuristic algorithms to find better network structures. Recently, several algorithms have been proposed to estimate optimal Bayesian network structure, but the number of genes included in the network is limited less than 30 or so. In order to apply Bayesian network approach to drug target gene discovery, we need to consider gene networks with several hundreds of genes. Therefore we need to develop more efficient algorithms to learn Bayesian network structure based on observed data. In this paper we propose an efficient structural learning algorithm for Bayesian networks by extending K2 algorithm that is one of the standard learning algorithms in Bayesian networks. We conduct Monte Carlo simulations to examine the effectiveness of the proposed algorithm by comparing with greedy hill-climbing algorithm. We also show the application of yeast gene network estimation based on the proposed algorithm.

### h. Clustering samples characterized by time course gene expression profiles using the mixture of state space models

**Osamu Hirose, Ryo Yoshida[5], Rui Yamaguchi, Seiya Imoto, Tomoyuki Higuchi[5], Satoru Miyano: [5]Institute of Statistical Mathematics**

We propose a novel method to classify samples where each sample is characterized by a time course gene expression profile. By exploiting the mixture of state space model, the proposed method addresses the following tasks: (1) clustering samples according to temporal patterns of gene expressions, (2) automatic detection of genes that discriminate identified clusters, (3) estimation of a restricted autoregressive coefficient for each cluster. We demonstrate the proposed method along with the cluster analysis of 53 multiple sclerosis patients under recombinant interferon βtherapy with the longitudinal time course expression profiles.

### i. Identification of activated transcription factors from microarray gene expression data of Kampo medicine-treated mice

**Rui Yamaguchi, Masahiro Yamamoto[6], Seiya Imoto, Masao Nagasaki, Ryo Yoshida[5], Kenji Tsuiji[6], Atsushi Ishige[6], Hiroaki Asou[7], Kenji Watanabe[2], Satoru Miyano: [6]Keio University School of Medicine, [7]Tokyo Metropolitan Institute of Gerontology**

We propose an approach to identify activated transcription factors from gene expression data using a statistical test. Applying the method, we can obtain a synoptic map of transcription factor activities which helps us to easily grasp the system's behavior. As a real data analysis, we use a case-control experiment data of mice treated by a drug of Kampo medicine remedying degraded myelin sheath of nerves in central nervous system. Kampo medicine is Japanese traditional herbal medicine. Since the drug is not a single chemical compound but extracts of multiple medicinal herb, the effector sites are possibly multiple. Thus it is hard to understand the action mechanism and the system's behavior by investigating only few highly expressed individual genes. Our method gives summary for the system's behavior with various functional annotations, e.g. TFAs and gene ontology, and thus offer clues to understand it in more holistic manner.

### j. AYUMS: an algorithm for completely automatic quantitation based on LC-MS/MS proteome data and its application to the analysis of signal transduction

**Ayumu Saito, Masao Nagasaki, Masaaki Oyama, Hiroko Kozuka-Hata, Kentaro Semba, Sumio Sugano, Tadashi Yamamoto, Satoru Miyano**

Comprehensive description of the behavior of cellular components in a quantitative manner is essential for systematic understanding of biological events. Recent LC-MS/MS (tandem mass spectrometry coupled with liquid chromatography) technology, in combination with the SILAC (Stable Isotope Labeling by Amino acids in Cell culture) method, has enabled us to make relative quantitation at the proteome level. The recent report by Blagoev *et al*. (Nat. Biotechnol., 22, 1139-1145, 2004) indicated that this method was also applicable for the time-course analysis of cellular signalling events. Relative quatitation can easily be performed by calculating the ratio of peak intensities corresponding to differentially labeled peptides in the MS spectrum. As currently available software requires some GUI applications and is time-consuming, it is not suitable for processing large-scale proteome data.

To resolve this difficulty, we developed an algorithm that automatically detects the peaks in each spectrum. Using this algorithm, we developed a software tool named AYUMS that automatically identifies the peaks corresponding to differentially labeled peptides, compares these peaks, calculates each of the peak ratios in mixed samples, and integrates them into one data sheet. This software has enabled us to dramatically save time for generation of the final report.

AYUMS is a useful software tool for comprehensive quantitation of the proteome data generated by LC-MS/MS analysis. This software was developed using Java and runs on Linux, Windows, and Mac OS X. Please contact ayums @ims.u-tokyo.ac.jp if you are interested in the application. The project web page is http:// www.csml.org/ayums/

### k. Modeling gene expression regulatory networks with the sparse vector autoregressive model

**Andre Fujita, Joao R Sato[8], Humberto M. Garay-Malpartida[8], Rui Yamaguchi, Satoru Miyano, Mari C. Sogayar[8], Carlos E Ferreira[8]: [8]University of São Paulo**

To understand the molecular mechanisms underlying important biological processes, a detailed description of the gene products networks involved is required. In order to define and understand such molecular networks, some statistical methods are proposed in the literature to estimate gene regulatory networks from time-

series microarray data. However, several problems still need to be overcome. Firstly, information flow need to be inferred, in addition to the correlation between genes. Secondly, we usually try to identify large networks from a large number of genes (parameters) originating from a smaller number of microarray experiments (samples). Due to this situation, which is rather frequent in Bioinformatics, it is difficult to perform statistical tests using methods that model large genegene networks. In addition, most of the models are based on dimension reduction using clustering techniques, therefore, the resulting network is not a gene-gene network but a module-module network. Here, we present the Sparse Vector Autoregressive model as a solution to these problems. We have applied the Sparse Vector Autoregressive model to estimate gene regulatory networks based on gene expression profiles obtained from time-series microarray experiments. Through extensive simulations, by applying the SVAR method to artificial regulatory networks, we show that SVAR can infer true positive edges even under conditions in which the number of samples is smaller than the number of genes. Moreover, it is possible to control for false positives, a significant advantage when compared to other methods described in the literature, which are based on ranks or score functions. By applying SVAR to actual HeLa cell cycle gene expression data, we were able to identify well known transcription factor targets. The proposed SVAR method is able to model gene regulatory networks in frequent situations in which the number of samples is lower than the number of genes, making it possible to naturally infer partial Granger causalities without any a priori information. In addition, we present a statistical test to control the false discovery rate, which was not previously possible using other gene regulatory network models.

## 2. Statistical and Computational Knowledge Discovery

### a. Statistical absolute evaluation of gene ontology terms with gene expression datas

**Pramod K. Gupta, Ryo Yoshida[5], Seiya Imoto, Rui Yamaguchi, Satoru Miyano**

We propose a new testing procedure for the automatic ontological analysis of gene expression data. The objective of the ontological analysis is to retrieve some functional annotations, e. g. Gene Ontology terms, relevant to underlying cellular mechanisms behind the gene expression profiles, and currently, a large number of tools have been developed for this purpose. The most existing tools implement the same approach that exploits rank statistics of the genes which are ordered by the strength of statistical evidences, e.g. p-values computed by testing hypotheses at the individual gene level. However, such an approach often causes the serious false discovery. Particularly, one of the most crucial drawbacks is that the rank-based approaches wrongly judge the ontology term as statistically significant although all of the genes annotated by the ontology term are irrelevant to the underlying cellular mechanisms. In this paper, we first point out some drawbacks of the rank-based approaches from the statistical point of view, and then, propose a new testing procedure in order to overcome the drawbacks. The method that we propose has the theoretical basis on the statistical meta-analysis, and the hypothesis to be tested is suitably stated for the problem of the ontological analysis. We perform Monte Carlo experiments for highlighting the disadvantages of the rank-based approach and the advantages of the proposed method. Finally, we demonstrate the applicability of the proposed method along with the ontological analysis of the gene expression data of human diabetes.

### b. Computational discovery of aberrant splice variations with genome-wide exon expression profiles

**Ryo Yoshida[5], Kazuyuki Numata, Seiya Imoto, Masao Nagasaki, Atsushi Doi[3], Kazuko Ueno, Satoru Miyano**

Alternative splicing plays a prominent role in eukaryotic gene regulations that allow a single gene to generate the multiple mRNA products. The recent advent of GeneChip Human Exon 1.0 ST Array enables us to measure the exon expression profiles of human cells on a genome-wide scale. With this advent, analysis of functional gene regulation could be extended to detect not only differentially expressed genes, but also specific splicing events that occur in target cells, but not in normal controls. We address some statistical issues for the identification of biomarker splice variations with exon expression data. The proposed method involves the following steps: (1) Whole transcript analysis with the nonparametric analysis of variance (ANOVA) to identify potential biomarkers that present specific splice variations. (2) Meta-analysis for discriminating non-specific splice variations that are caused by clinical heterogeneity in the collected samples. In the analysis of human cells, controlling non-specific splicing factors is essential for success in the detection of

biomarker splice variations because splice patterns are possibly affected by inter-individual differences in the collected samples. We demonstrate its utility and perform a whole transcript analysis of exon expression profiles of colorectal carcinoma.

### c. Performance improvement in protein N-myristoyl classification by BONSAI with insignificant indexing symbol

**Manabu Sugii[1], Ryo Okada[1], Hiroshi Matsuno[1], Satoru Miyano**

Many N-myristoylated proteins play key roles in regulating cellular structure and function. In the previous study, we have applied the machine learning system BONSAI to predict patterns based on which positive and negative examples could be classified. Although BONSAI has helped establish 2 interesting rules regarding the requirements for N-myristoylation, the accuracy rates of these rules are not satisfactory. This paper suggests an enhancement of BONSAI by introducing an "insignificant indexing symbol"and demonstrates the efficiency of this enhancement by showing an improvement in the accuracy rates. We further examine the performance of this enhanced BONSAI by comparing the results of classification obtained the proposed method and an existing public method for the same sets of positive and negative examples.

## Publications


1. Affara, M., Dunmore, B., Savoie, C.J., Imoto, S., Tamada, Y., Araki, H., Charnock-Jones, D.S., Miyano, S., Print, C. Understanding endothelial cell apoptosis: What can the transcriptome glycome and proteome reveal? Philosophical Transactions of Royal Society. 362 (1484): 1469-1487, 2007.
2. Akutsu, T., Bannai, H., Miyano, S., Ott, S. On the complexity of deriving position specific score matrices from positive and negative sequences. Discrete Applied Mathematics. 155: 676-685, 2007.
3. Fujita, A., Sato, J.R., Garay-Malpartida, H. M., Sogayar, M.C., Ferreira, C.E., Miyano, S. Modeling nonlinear gene regulatory networks from time series gene expression data. J. Bioinformatics and Computational Biology. In press.
4. Fujita, A., Sato, J.R., Garay-Malpartida, H. M., Yamaguchi, R., Miyano, S., Sogayar, M. C., Ferreira, C.E. Modeling gene expression regulatory networks with the sparse vector autoregressive model. BMC Systems Biology 2007, 1: 39 (doi:10.1186/1752-0509-1-39)
5. Gupta, P.K., Yoshida, R., Imoto, S., Yamaguchi, R., Miyano, S. Statistical absolute evaluation of gene ontology terms with gene expression data. Lecture Notes in Bioinformatics. 4463: 146-157, 2007.
6. Hirose, O., Yoshida, R., Imoto, S., Yamaguchi, R., Higuchi, T., Charnock-Jones, D.S., Print, C., Miyano, S. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. Bioinformatics. In press.
7. Hirose, O., Yoshida, R., Yamaguchi, R., Imoto, S., Higuchi, T., Miyano, S. Clustering with time course gene expression profiles and the mixture of state space models. Genome Informatics. 18: 258-266, 2007.
8. Imoto, S. Knowledge discovery of causal relations among genes from microarray gene expression data, Journal of Japan Statistical Society. 37 (1): 55-70, 2007.
9. Imoto, S., Miyano, S. Bayesian network approach to estimate gene networks. A. Mittal, A. Kassim and T. Tan (Eds.), Bayesian Network Technologies : Applications and Graphical Models, Idea Group Publishers, USA. 269-299, 2007.
10. Imoto, S., Tamada, Y., Savoie, C.J., Miyano, S., Analysis of gene networks for drug target discovery and validation. Methods in Molecular Biology. 360: 33-56, 2007.
11. Jeong, E., Nagasaki, M., Miyano, S. Conversion from BioPAX to CSO for system dynamics and visualization of biological pathway. Genome Informatics. 18: 225-236, 2007.
12. Jeong, E., Nagasai, M., Saito, A., Miyano, S. Cell System Ontology: Representation for modeling, visualizing, and simulating biological pathways. In Silico Biology 7, 0055, 2007.
13. Kojima, K., Nagasaki, M., Jeong, E., Kato, M., Miyano, S. An efficient grid layout algorithm for biological networks utilizing various biological attributes. BMC Bioinformatics. 8: 76, 2007.
14. Li, C., Ge, Q.-W., Nakata, M., Matsuno, H., Miyano, S. Modeling and simulation of signal transductions in an apoptosis pathway by using timed Petri nets. J. Biosciences. 32 (1): 113-125, 2007.
15. Miyano, S., DeLisi, C., Holzhütter, H.-G., Kanehisa, M. (Eds.). Genome Informatics. 18,

2007.
16. Numata, K., Imoto, S., Miyano, S.A structure learning algorithm for inference of gene networks from microarray gene expression data using Bayesian networks. Proc. IEEE 7th International Symposium on Bioinformatics & Bioengineering. 1280-1284, 2007. (BIBE2007: Refereed conference; Digital Object Identifier 10.1109/BIBE.2007.4375731)
17. Saito, A., Nagasaki, M., Oyama, M., Kozuka-Hata, H., Semba, K., Sugano, S., Yamamoto, T., Miyano, S. AYUMS: an algorithm for completely automatic quantitation based on LC-MS/MS proteome data and its application to the analysis of signal transduction. BMC Bioinformatics. 8: 15, 2007.
18. Shimamura, T., Yamaguchi, R., Imoto, S., Miyano, S. Weighted lasso in graphical Gaussian modeling for large gene network estimation based on microarray data. Genome Informatics. 19: 142-153, 2007.
19. Sugii, M., Okada, R., Matsuno, H., Miyano, S. Performance improvement in protein N-myristoyl classification by BONSAI with insignificant indexing symbol. Genome Informatics. 18: 277-286, 2007.
20. Termier, A., Tamada, Y., Numata, K., Imoto, S., Washio, T., Higuchi, T., DIGDAG, a first algorithm to mine closed frequent embedded sub-DAGs. Proc. 5th International Workshop on Mining and Learning with Graphs, in press, 2007.
21. Yamaguchi, R., Yamamoto, M., Imoto, S., Nagasaki, M., Yoshida, R., Tsujii, K., Ishiga, A., Asou, H., Watanabe, K., Miyano, S. Identification of activated transcription factors from microarray gene expression data of Kampo-medicine treated mice. Genome Informatics. 18: 119-129, 2007.
22. Yamaguchi, R., Yoshida, R., Imoto, S., Higuchi, T., Miyano, S. Finding module-based gene networks with state-space models? Mining high-dimensional and short time-course gene expression data. IEEE Signal Processing Magazine, 24 (1): 37-46, 2007.
23. Yoshida, R., Numata, K., Imoto, S., Nagasaki, M., Doi, A., Ueno, K., Miyano, S. Computational discovery of aberrant splice variations with genome-wide exon expression profiles. Proc. IEEE 7th International Symposium on Bioinformatics & Bioengineering. 715-722, 2007. (IEEE BIBE2007: Refereed conference; Digital Object Identifier 10.1109/BIBE.2007.4375639)

*Human Genome Center*

# Laboratory of Molecular Medicine
# Laboratory of Genome Technology
# Division of Advanced Clinical Proteomics

## 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
## 先端臨床プロテオミクス共同研究ユニット


| | | | | | |
|---|---|---|---|---|---|
| Professor | Yusuke Nakamura, M.D., Ph.D. | 教　授 | 医学博士 | 中　村　祐　輔 |
| Associate Professor | Toyomasa Katagiri, Ph.D. | 准教授 | 医学博士 | 片　桐　豊　雅 |
| | Hidewaki Nakagawa, M.D., Ph.D. | 准教授 | 医学博士 | 中　川　英　刀 |
| Assistant Professor | Ryuji Hamamoto, Ph.D. | 助　教 | 理学博士 | 浜　本　隆　二 |
| | Koichi Matsuda, M.D., Ph.D. | 助　教 | 医学博士 | 松　田　浩　一 |
| | Hitoshi Zembutsu, M.D., Ph.D. | 助　教 | 医学博士 | 前　佛　　均 |
| Project Associate Professor | Yataro Daigo, M.D., Ph.D. | 特任准教授 | 医学博士 | 醍　醐　弥太郎 |



*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


**Toyomasa Katagiri, Yataro Daigo, Hidewaki Nakagawa, Hitoshi Zembutsu, Koichi Matsuda, Ryo Takata, Mitsugu Kanehira, Koji Takahashi, Atsushi Takano, Nobuhisa Ishikawa, Tatsuya Kato, Satoshi Hayama, Chie Suzuki, Akira Togashi, Kazuhito Morioka, Tomohide Kidokoro, Chizu Tanikawa, Asahi Hishida, Miki Akiyama, Sachiko Dobashi, Meng-Lay Lin, Jae-Hyun Park, Tomomi Ueki, Yosuke Harada, Chikako Fukukawa, Toshihiko Nishidate, Koji Ueda, Nguyen Minh-Hue, Yuria Mano, Masaya Taniwaki, Ryohei Nishino, Daizaburo Hirata, Takumi Yamabuki, Nagato Sato, Junkichi Koinuma, Banu Turel, Kenji Tamura, Masayo Hosokawa, Su-Youn Chung, Motohide Uemura, Akio Takehara, Arata Shimo, Fabio Pittella Silva and Yusuke Nakamura**


#### (1) Lung cancer

**MAPJD (Myc-associated protein with JmjC domain)**

Through genome-wide expression profile analysis for non-small cell lung carcinomas (NSCLCs), we found over-expression of a *MAPJD* (Myc-associated protein with JmjC domain) gene in the great majority of NSCLC cases. Induction of exogenous expression of MAPJD into NIH3T3 cells conferred growth-promoting activity. Concordantly, *in vitro* suppression of MAPJD expression with siRNA effectively suppressed growth of NSCLC cells in which MAPJD was over-expressed. We found four candidate MAPJD-target genes, *SBNO1, TGFBRAP1, RIOK1*, and *RASGEF1A*, which were the most significantly induced by exogenous MAPJD expression. Through interaction with MYC protein, MAPJD transactivates a set of genes including kinases and cell signal transducers that are possibly related to proliferation of lung cancer cells. As our data imply that MAPJD is a novel member of the MYC transcriptional complex and its activation is a common feature of lung-cancer, selective suppression of this pathway could be a promising therapeutic target for treatment of lung cancers.

### CDCA8 (cell-division-associated 8)

We found in the great majority of lung-cancer samples co-transactivation of cell-division-associated 8 (*CDCA8*) and aurora kinase B (*AURKB*) that were considered to be components of the vertebrate chromosomal passenger complex. Immunohistochemical analysis of lung-cancer tissue microarrays demonstrated that over-expression of CDCA8 and AURKB was significantly associated with poor prognosis of lung-cancer patients. AURKB directly phosphorylated CDCA8 at Ser-154, Ser-219, Ser-275, and Thr-278, and appeared to stabilize CDCA8 protein in cancer cells. Suppression of CDCA8 expression with siRNA against *CDCA8* significantly suppressed the growth of lung cancer cells. In addition, functional inhibition of interaction between CDCA8 and AURKB by a cell-permeable peptide corresponding to 20 amino-acid sequence of a part of CDCA8 (11R-CDCA $8_{261-280}$), which included two phosphorylation sites by AURKB, significantly reduced phosphorylation of CDCA8 and resulted in growth suppression of lung cancer cells. Our data imply that selective suppression of the CDCA8-AURKB pathway could be a promising therapeutic strategy for treatment of lung cancer patients.

### HJURP (Holliday Junction-Recognizing Protein)

We additionally identified a novel gene encoding *HJURP* (Holliday Junction-Recognizing Protein) whose activation appeared to play a pivotal role in immortality of cancer cells. HJURP was shown to be a possible downstream target of ATM signaling and increased by DNA double-strand breaks (DSBs). HJURP is involved in the homologous recombination (HR) pathway in the DSBs repair process, interacting with hMSH5, or a member of a MRN protein complex NBS1. HJURP formed nuclear foci in cells at S-phase or in those after having DNA damage. *In vitro* assays implied that HJURP directly bound to Holliday Junction and rDNA array. Treatment of cancer cells with siRNA against HJURP caused abnormal chromosomal fusions, and led to genomic instability and senescence. In addition, HJURP over-expression was observed in the majority of lung cancers and associated with poorer prognosis. We suggest that HJURP is an indispensable factor for chromosomal stability in immortalized cancer cells and could be a novel therapeutic target for development of anti-cancer drugs.

### FGFR1OP (fibroblast growth factor receptor 1 oncogene partner)

We found an elevated expression of FGFR1OP (fibroblast growth factor receptor 1 oncogene partner) in the great majority of lung cancers. Immunohistochemical staining using tumor tissue microarrays consisting of 372 archived non-small cell lung cancer (NSCLC) specimens revealed positive staining of FGFR1OP in 334 (89.8%) of 372 NSCLCs. We also found that the high level of FGFR1OP expression was significantly associated with shorter tumor-specific survival times ($P < 0.0001$ by log-rank test). Moreover multivariate analysis determined that FGFR1OP was an independent prognostic factor for surgically-treated NSCLC patients ($P < 0.0001$). Treatment of lung cancer cells, in which endogenous FGFR1OP was over-expressed, using FGFR1OP siRNA, suppressed its expression and resulted in inhibition of the cell growth. Furthermore, induction of FGFR1OP increased the cellular motility, invasion, and growth-promoting activity of mammalian cells. To investigate its function, we searched for FGFR1 OP-interacting proteins in lung cancer cells and identified the ABL1 (Abelson murine leukemia viral oncogene homolog 1) and WRNIP1 (Werner helicase interacting protein 1), which was known to be involved in cell cycle progression. ABL1 appeared to suppress DNA synthesis by directly phosphorylating tyrosine resides of WRNIP1, while FGFR1OP significantly reduced ABL1-dependent phosphorylation of WRNIP1 and resulted in the promotion of DNA synthesis

of cancer cells. Since our data imply that FGFR1 OP is likely to play a significant role in lung cancer growth and progression, FGFR1OP should be useful as a prognostic biomarker and probably as a therapeutic target for lung cancer.

### (2) Breast Cancer

#### MELK (maternal embryonic leucine-zipper kinase)

Cancer therapy directing at specific molecular targets in signaling pathways of cancer cells such as Tamoxifen, aromatase inhibitors and trastuzumab has been proven its usefulness for treatment of advanced breast cancers. However, increases of the risk of endometrial cancer by long-term tamoxifen administration as well as those of bone fracture due to osteoporosis in postmenopausal women with aromatase inhibitor prescription are recognized as their side effects. Due to the emergence of these side effects and also drug resistance, it is necessary to search novel targets for molecularly-orientated drugs on the basis of well-characterized mechanisms of action. Using the accurate genome-wide expression profiles of breast cancers, we found *maternal embryonic leucine-zipper kinase* (*MELK*) that was significantly overexpressed in the great majority of breast cancer cells. To assess a possible role of *MELK* in mammary carcinogenesis, we knocked down the expression of endogenous *MELK* in breast cancer cell-lines by means of the mammalian vector-based RNA interference (RNAi) technique. Furthermore, we identified a long isoform of Bcl-G (Bcl-$G_L$), a pro-apoptotic member of Bcl-2 family, as a possible substrate(s) for the MELK kinase by pull-down assay with wild-type- and kinase-dead-MELK recombinant proteins. Finally, we performed TUNEL assay and FACS analysis to measure the proportions of sub-G1 population to investigate MELK is involved in apoptosis cascade through the Bcl-$G_L$-related pathway. The multiple human tissues- and cancer cell lines-northern blot analyses demonstrated that *MELK* was overexpressed at a significantly high level in a great majority of breast cancers and cancer cell lines, but not expressed in normal vital organs (heart, liver, lung, hidney). Suppression of *MELK* expression with small-interfering RNA significantly inhibited growth of human breast cancer cells. We also found that MELK protein physically interacted with Bcl-$G_L$ protein, a pro-apoptotic member of the Bcl-2 family through its N-terminal region. Subsequent immunocomplex kinase assay showed Bcl-$G_L$ was specifically phosphorylated by MELK *in vitro*. TUNEL assay and FACS analysis revealed that overexpression of WT-MELK suppressed Bcl-$G_L$-induced apoptosis, while that of D150A-MELK did not. Our findings suggest that kinase activity of MELK is likely to be involved in mammary carcinogenesis through inhibition of pro-apoptotic function of Bcl-$G_L$. The kinase activity of MELK should be a promising target for development of molecular-targeting therapy for patients with breast cancers.

#### KIF2C/MCAK (Kinesin family member 2C/Mitotic centromere-associated kinesin)

We also demonstrated functional significance of *KIF2C/MCAK* (Kinesin family member 2C/Mitotic centromere-associated kinesin) in growth of breast cancer cells. Northern-blot and immunohistochemical analyses confirmed KIF2C/MCAK overexpression in breast cancer cells, and indicated its undetectable level of expression in normal human tissues except testis, suggesting KIF2C/MCAK to be a cancer-testis antigen. Western-blot analysis using breast cancer cell-lines revealed a significant increase of endogenous KIF2C/MCAK protein level and its phosphorylation in $G_2$/M phase. Treatment of breast cancer cells with small-interfering RNAs (siRNAs) against KIF2C/MCAK effectively suppressed KIF2C/MCAK expression and inhibited the growth of breast cancer cell-lines, T47D and HBC5 cells. We also demonstrated that KIF2C/MCAK expression was significantly suppressed by ectopic introduction of p53. These findings suggest that overexpression of KIF2C/MCAK might be involved in breast carcinogenesis and is a promising therapeutic target for breast cancers.

### (3) Bladder cancer

#### MPHOSPH1 (M-phase phosphoprotein 1)

To disclose molecular mechanism of bladder cancer, the second most common genitourinary tumor, we had previously performed genome-wide expression profile analysis of 26 bladder cancers by means of cDNA microarray representing 27,648 genes. Among genes that were significantly up-regulated in the majority of bladder cancers, we here report identification of *MPHOSPH1* (*M-phase phosphoprotein 1*) as a candidate molecule for drug development for bladder cancer. Northern blot analyses using mRNAs of normal human organs and cancer cell-lines indicated this molecule to be a novel cancer / testis antigen . Introduction of *MPHOSPH1* into NIH3T3 cells significantly enhanced cell growth at *in vitro* and *in vivo* conditions. We subsequently found an interaction be-

tween MPHOSPH1 and PRC1 (Protein Regulator of Cytokinesis 1), which was also up-regulated in bladder cancer cells. Immunocytochemical analysis revealed co-localization of endogenous MPHOSPH1 and PRC1 proteins in bladder cancer cells. Interestingly, knockdown of either of MPHOSPH1 or PRC1 expression with specific siRNAs caused significant increase of multi-nuclear cells and subsequent cell death of bladder cancer cells. Our results imply that the MPHOSPH1/PRC1 complex is likely to play a crucial role in bladder carcinogenesis and that inhibition of the MPHOSPH1/PRC1 expression or their interaction should be novel therapeutic targets for bladder cancers.

### DEPDC1 (DEP domain containing 1)

We also found a novel gene, *DEPDC1* (*DEP domain containing 1*), whose over-expression was confirmed in bladder cancers by northern-blot and immunohistochemical analyses. Immunocytochemical staining analysis detected strong staining of endogenous DEPDC1 protein in the nucleus of bladder cancer cells. Since *DEPDC1* expression was hardly detectable in any of 24 normal human tissues we examined except the testis, we considered this gene-product to be a novel cancer/testis antigen. Suppression of *DEPDC1* expression with small-interfering RNA (siRNA) significantly inhibited growth of bladder cancer cells. Taken together, these findings suggest that *DEPDC1* might play an essential role in the growth of bladder cancer cells, and would be a promising molecular-target for novel therapeutic drugs or cancer peptide-vaccine to bladder cancers.

## (4) Pancreatic cancer

### KIAA0101

Pancreatic ductal adenocarcinoma (PDAC) shows the worst mortality among the common malignancies, with a 5-year survival rate of only 4%, and the majority of PDAC patients are diagnosed at an advanced stage, in which no effective therapy is available at present. Although the proportion of curable cases is still not so high, surgical resection of early-staged PDACs is the only way to cure the disease. Hence, establishment of a screening strategy to detect early-staged PDACs by novel serological markers is urgently required, and development of novel molecular therapies for PDAC treatment is also eagerly expected. To isolate novel diagnostic markers and therapeutic targets for pancreatic cancer, we earlier performed expression profile analysis of pancreatic cancer cells using a genome-wide cDNA microarray combined with laser microdissection. Among dozens of transactivated genes in pancreatic cancer cells, this study focused on *KIAA0101* whose overexpression in pancreatic cancer cells was validated by immunohitochemical analysis. KIAA0101 was previously identified as $p15^{PAF}$ (PCNA-associated factor) to bind with PCNA (proliferating cell nuclear antigen), however its function remains unknown. To investigate for the biological significance of KIAA0101 overexpression in cancer cells, we knocked down *KIAA0101* by siRNA in pancreatic cancer cells, and found that the reduced expression by siRNA caused drastic attenuation of their proliferation as well as significant decrease in DNA replication rate. Concordantly, exogenous over-expression of KIAA0101 enhanced cancer cell growth and NIH3T3-derivative cells expressing KIAA0101 revealed *in vivo* tumor formation, implying its growth promoting and oncogenic property. We also demonstrated that the expression of KIAA0101 was regulated tightly by p53-p21 pathway. To consider the KIAA0101/PCNA interaction as a therapeutic target, we designed the cell-permeable 20-amino-acid dominant-negative peptide and found that it could effectively inhibit the KIAA0101/PCNA interaction and resulted in the significant growth suppression of cancer cells. Our results clearly implicated that suppression of the KIAA0101 and PCNA oncogenic activity, or the inhibition of KIAA0101/PCNA interaction is likely to be a promising strategy to develop novel cancer therapeutic drugs.

### GABRP (GABA receptor π subunit)

GABA (γ-amino-butyric acid) functions primarily as an inhibitory neurotransmitter in the mature central nervous system, and GABA/GABA receptors are also present in non-neural tissues, including cancer, but their precise function in non-neuronal or cancerous cells is poorly defined so far. We identified over-expression of GABA receptor π subunit (GABRP) in pancreatic ductal adenocarcinoma (PDAC) cells. We also found the expression of this peripheral-type $GABA_A$ receptor subunit, GABRP, in some adult human organs. Knockdown of endogenous *GABRP* expression in PDAC cells by siRNA attenuated PDAC cell growth, suggesting its essential role in PDAC cell viability. Notably, addition of GABA into the cell-culture medium promoted the proliferation of GABRP-expressing PDAC cells, but not GABRP-negative cells, and $GABA_A$ receptor antagonists inhibited this growth-promoting effect by GABA. The HEK293 cells constitutively expressing exogenous

GABRP revealed the growth-promoting effect by GABA treatment. Furthermore, GABA treatment in GABRP-positive cells increased intracellular $Ca^{2+}$ level and activated MAPK (mitogen-activated protein kinase) cascade. Clinical PDAC tissues contained a higher level of GABA ligand than normal pancreas tissues due to the up-regulation of GAD1 (glutamate decarboxylase 1) expression, suggesting their autocrine/paracrine growth-promoting effect in PDACs. These findings imply that GABA ligand and GABRP could play important roles in PDAC development and progression, and this pathway can be a promising molecular target for development of new therapeutic strategies for PDAC.

### (4) Prostate cancer

One of the most critical issues in prostate cancer clinic is emerging hormone-refractory prostate cancers (HRPCs) and their management. Prostate cancer is usually androgen-dependent and responds well to androgen ablation therapy. However, at a certain stage they eventually acquire androgen-independent phenotype and show very poor response to any anti-cancer therapies that are presently available. To characterize the molecular features of clinical HRPCs, we analyzed gene-expression profiles of 25 clinical HRPCs and 10 hormone-sensitive prostate cancers (HSPCs) by genome-wide cDNA microarrays combining with laser microbeam microdissection. An unsupervised clustering analysis clearly distinguished expression patterns of HRPC cells from those of HSPC cells. In addition, primary and metastatic HRPCs from individual patients were closely clustered regardless to metastatic organs. A supervised clustering analysis identified 36 up-regulated genes and 70 down-regulated genes in HRPCs, compared with HSPCs ($P < 0.0001$, gap $> 1.5$). We observed over-expression of *AR, ANLN*, and *SNRPE*, and down-regulation of *NR4A1*, *CYP27A1*, and *HLA-A* antigen in HRPC progression. Such genes were considered to be related to the androgen-independent and more aggressive phenotype of HRPCs, and in fact, knockdown of some over-expressing genes by siRNA resulted in drastic attenuation of prostate cancer cell growth. Our precise microarray analysis of HRPC cells should provide useful information to understand the molecular mechanism of HRPC progression as well as to identify molecular targets for development of treatment of HRPCs.

#### SRD5A3

We then identified one novel gene, *SRD5A2L*, encoding a putative 5α-steroid reductase that produces the most potent androgen DHT (5α-dihydrotestosterone) from testosterone. LC-MS/MS analysis following *in vitro* 5α-steroid reductase reaction validated its ability to produce DHT from testosterone as similar to type-1 5α-steroid reductase. Since two types of 5α-steroid reductases were previously reported, we termed this novel 5α-steroid reductase to be *SRD5A3* (type-3 5α-steroid reductase). RT-PCR and northern blot analyses confirmed its over-expression in HRPC cells, and indicated no or little expression in normal adult organs. Knockdown of *SRD5A3* expression by siRNA in prostate cancer cells resulted in significant decrease of DHT production and drastic reduction of their cell viability. These findings implicate that a novel 5α-steroid reductase, SRD5A3, is associated with DHT production and maintenance of the AR pathway activation in HRPC cells, and that this enzymatic activity should be a promising molecular target for prostate cancer therapy.

### (5) Chemosensitivity

#### Bladder Cancer

To predict the efficacy of the M-VAC neoadjuvant chemotherapy for invasive bladder cancers, we previously established the method to calculate the prediction score on the basis of expression profiles of 14 predictive genes. This scoring system had clearly distinguished the responder group from the non-responder group. To further validate the clinical significance of the system, we applied it to 22 additional cases of bladder cancer patients and found that the scoring system correctly predicted clinical response for 19 of the 22 test cases. The group of patients with positive predictive scores had significantly longer survival than that with negative scores. When we compared our results with the previous report describing the prognosis of the patients with cystectomy alone, the results imply that patients with positive scores are likely to have benefit by having M-VAC neoadjuvant chemotherapy, but that the chemotherapy would shorten lives of patients with negative scores. We are confident that our prediction system to M-VAC therapy should provide opportunities for achieving better prognosis and improving quality of life of patients.

#### Philadelphia chromosome-positive acute lymphoblastic leukemia (Ph+ALL)

Philadelphia chromosome-positive acute lymphoblastic leukemia (Ph+ALL) reveals very poor prognosis due to high incidence of relapse

when treated with standard chemotherapy. Although more than 96 % of patients with Ph+ ALL achieved complete remission (CR) with imatinib-combined chemotherapy in a phase II clinical trial conducted by the Japan Adult Leukemia Study Group (JALSG), 26 % of them experienced hematological relapse in a short time after achievement of CR. In this study, to establish a prediction system for risk of relapse, we analyzed gene expression profiles of 23 bone marrow samples from patients with Ph+ALL using cDNA microarray consisting of 27,648 cDNA sequences. Using the 19 randomly-selected test cases, we identified 16 genes that were expressed significantly differently between patients with (n=8) and without (n=11) continuous response; from the list of 16 genes, we selected the 6 "predictive" genes and constructed a numerical scoring system by which the two groups were clearly separated, with positive scores for the former and the negative scores for the latter. Scoring of 4 cases that were reserved from the original 23 cases predicted correctly their clinical responses. In addition, three cases whose BCR-Abl transcript levels failed to reduce sufficiently after induction therapy, also revealed negative scores. We also developed a quantitative reverse transcription-PCR -based prediction system that could be feasible for routine clinical use. Our results suggest that achievement of continuous response with imatinib-combined chemotherapy can be predicted by expression patterns in this set of genes, leading to achievement of "personalized therapy" for treatment of this disease.

### (6) Biomarker

### DKK1 (Dickkopf-1)

Gene-expression profile analysis of lung and esophageal carcinomas revealed that Dickkopf-1 (DKK1) was highly transactivated in the great majority of lung cancers and esophageal squamous-cell carcinomas (ESCCs). Immunohistochemical staining using tumor tissue microarrays consisting of 279 archived non-small cell lung cancers (NSCLCs) and 280 ESCC specimens demonstrated that a high level of DKK1 expression was associated with poor prognosis of patients with NSCLC as well as ESCC, and multivariate analysis confirmed its independent prognostic value for NSCLC. In addition, we identified that exogenous expression of DKK1 increased the migratory activity of mammalian cells, suggesting that DKK1 may play a significant role in progression of human cancer. We established an ELISA system to measure serum levels of DKK1 and found that serum DKK1 levels were significantly higher in lung and esophageal cancer patients than in healthy controls. The proportion of the DKK1-positive cases was 126 (70.0%) of 180 NSCLC, 59 (69.4%) of 85 SCLC, and 51 (63.0%) of 81 ESCC patients, while only 10 (4.8%) of 207 healthy volunteers were falsely diagnosed as positive. A combined ELISA assays for both DKK1 and CEA increased sensitivity, and classified 82.2% of the NSCLC patients as positive while only 7.7% of healthy volunteers were falsely diagnosed to be positive. The use of both DKK1 and ProGRP increased sensitivity to detect SCLCs up to 89.4%, while false positive rate in healthy donors were only 6.3%. Our data imply that DKK1 should be useful as a novel diagnostic/prognostic biomarker in clinic and probably as a therapeutic target for lung and esophageal cancer.

### KIF4A

To identify molecules that might be useful as diagnostic/prognostic biomarkers and as targets for the development of new molecular therapies, we screened genes that were highly transactivated in a large proportion of 101 lung cancers by means of a cDNA microarray representing 27,648 genes. We found a gene encoding KIF4A, a kinesin family member 4A, as one of such candidates. Tumor-tissue microarray was applied to examine the expression of KIF4A protein and its clinicopathological significance in archival non-small cell lung cancer (NSCLC) samples from 357 patients. A role of KIF4A in cancer cell growth and/or survival was examined by small interfering RNA (siRNA) experiments. Cellular invasive activity of KIF4A on mammalian cells was examined using Matrigel assays. Immunohistochemical staining detected positive KIF4A staining in 127 (36%) of 357 NSCLCs and 19 (66 %) of 29 SCLCs examined. Positive immunostaining of KIF4A protein was associated with male gender ($P = 0.0287$), non-adenocarcinoma histology ($P=0.0097$), and shorter survival for patients with NSCLC ($P = 0.0005$), and multivariate analysis confirmed its independent prognostic value ($P=0.0012$). Treatment of lung cancer cells with siRNAs for *KIF4A* suppressed growth of the cancer cells. Furthermore, we found that induction of exogenous expression of KIF4A conferred cellular invasive activity on mammalian cells. These data strongly implied that targeting the KIF4A molecule might hold a promise for the development of anti-cancer drugs and cancer vaccines as well as a prognostic biomarker in clinic.

### LY6K (lymphocyte antigen 6 complex, locus K)

We revealed that a member of low molecular weight GPI-anchored molecule-like protein, lymphocyte antigen 6 complex, locus K (LY6K) was transactivated in the great majority of NSCLCs and ESCCs. Northern-blot analysis detected expression of *LY6K* gene only in testis among normal tissues examined. Immunohistochemical staining using tumor tissue microarrays consisting of 413 archived NSCLC and 271 ESCC specimens confirmed that a high level of LY6K expression was associated with poor prognosis of patients with NSCLC as well as ESCC, and multivariate analysis confirmed its independent prognostic value for NSCLC. In addition, treatment of NSCLC cells with small interfering RNAs against *LY6K* knocked-down its expression and resulted in growth suppression of the cancer cells. Serum levels of LY6K were significantly higher in NSCLC and ESCC patients than in healthy controls. The proportion of the serum LY6K-positive cases defined by our criteria was 38 (33.9%) of 112 NSCLC and 26 (32.1%) of 81 ESCC, while only 3 (4.1%) of 74 healthy volunteers were falsely diagnosed as positive. A combined assay using both LY6K and CEA increased sensitivity, and judged 64.7% of the lung adenocarcinoma patients as positive while 9.5% of healthy volunteers were falsely diagnosed to be positive. The use of both LY6K and CYFRA 21-1 increased sensitivity to detect lung squamous-cell carcinomas up to 70.4%, while false positive rate in healthy donors were only 6.8%. Our data imply LY6K to be a cancer-testis antigen and suggest that LY6K should be useful as a diagnostic/prognostic biomarker and probably as a therapeutic target for development of new molecular-targeted agents and/or immunotherapy of lung and esophageal cancers.

### Proteomics

The recent progress in various proteomic technologies allows us to screen serum biomarker including carbohydrate antigens. However, only a limited number of proteins could be detected by current conventional methods such as shotgun-proteomics, primarily because of the enormous concentration distribution of serum proteins and peptides. To circumvent this difficulty and isolate potential cancer-specific biomarkers for diagnosis and treatment, we established a new screening system consisting of the sequential steps of (1) immuno-depletion of 6 high abundant proteins, (2) targeted enrichment of glycoproteins by lectin column chromatography, and (3) the quantitative proteome analysis using $^{12}C_6$– or $^{13}C_6$–NBS ( 2-nitrobenzensulfenyl) stable isotope labeling followed by MALDI-QIT-TOF mass spectrometric analysis. Through this systematic analysis for five serum samples derived from patients with lung adenocarcinoma, we identified as candidate biomarkers 34 serum glycoproteins that revealed significant difference in α1,6-fucosylation level between lung cancer and healthy control, clearly demonstrating that the carbohydrate-focused proteomics could allow for the detection of serum components with cancer-specific features. In addition, we developed more simplified and practical technique, mass spectrometry-based glycan structure analysis and lectin blotting, in order to validate glycan structure of candidate biomarkers that could be applicable in clinical use. Our new glycoproteomic strategy will provide highly-sensitive and quantitative profiling of specific glycan structures on multiple proteins, which should be useful for serum biomarker discovery.

## 2. Common diseases

### (1) Cerebral infarction

**Michiaki Kubo[1,2,4], Jun Hata[1,2,4], Toshiharu Ninomiya[1,2], Koichi Matsuda[4], Koji Yonemoto[1], Toshiaki Nakano[2,3], Tomonaga Matsushita[2,4], Keiko Yamanaka-Yamazaki[4], Yozo Ohnishi[5], Susumu Saito[5], Takanori Kitazono[2], Setsuro Ibayashi[2], Katsuo Sueishi[3], Mitsuo Iida[2], Yusuke Nakamura[4], and Yutaka Kiyohara[1].: [1]Department of Environmental Medicine, [2]Department of Medicine and Clinical Science, [3]Pathophysiological and Experimental Pathology, Graduate School of Medical Sciences, Kyushu University, Fukuoka 812-8582, Japan. [4]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan. [5]Laboratory for Genotyping, SNP Research Center, the Institute of Physical and Chemical Research (RIKEN), Yokohama 230-0045, Japan.**

### PRKCH

Cerebral infarction is the most common type of stroke and often causes long-term disability. To investigate the genetic contribution to cerebral infarction, we conducted a case-control study using 52,608 gene-based tag-SNPs selected from JSNP database. Here we reported that one non-synonymous SNP in a member of protein kinase C (PKC) family, *PRKCH*, was significantly associated with lacunar infarction in two independent Japanese samples ($p = 3.2 \times 10^{-7}$, crude odds ratio of 1.39). This SNP was likely to

affect the PKC activity. Furthermore, a 14-year follow-up cohort study in Hisayama (Fukuoka, Japan) supported involvement of this SNP for the development of cerebral infarction ($p=0.03$, age- and sex-adjusted hazard ratio of 2.83). We also found that PKCη was mainly expressed in vascular endothelial cells and foamy macrophages in human atherosclerotic lesions, and its expression was enhanced as the lesion type progressed. Our results support a role for *PRKCH* in the pathogenesis of cerebral infarction.

### AGTRL1 (angiotensin receptor like-1)

Comparison of allele frequencies between 1,112 cases with brain infarction and age- and sex-matched control subjects of the same number found a SNP in the 5′-flanking region of angiotensin receptor like-1 (*AGTRL1*) gene (rs 9943582, -154G/A) to have a significant association with brain infarction (Odds ratio=1.30, 95 % confidence interval (CI)=1.14-1.47, $P=0.000066$). We also found the binding of Sp1 transcription factor to the region including the susceptible G allele, but not the non-susceptible A allele. Luciferase assay and RT-PCR analysis demonstrated that exogenously-introduced Sp1 induced transcription of *AGTRL1* and its ligand, apelin as well, indicating direct regulation of apelin/APJ pathway by Sp1. Furthermore, a 14-year follow-up cohort study in a Japanese community in Hisayama town, Japan revealed that the homozygote of the susceptible G allele of this particular SNP had significantly higher risk of brain infarction (Hazard ratio=2.00, 95 % CI =1.22-3.29, $P=0.006$). Our results indicate that the SNP in the *AGTRL1* gene is associated with the susceptibility to brain infarction.

### (2) Kawasaki disease

**Yoshihiro Onouchi[1], Tomohiko Gunji[1,2], Jane C. Burns[3], Chisato Shimizu[3], Jane W. Newburger[4], Mayumi Yashiro[5], Yoshikazu Nakamura[5], Hiroshi Yanagawa[6], Keiko Wakui[7], Yoshimitsu Fukushima[7], Fumio Kishi[8], Kunihiro Hamamoto[9], Masaru Terai[10], Yoshitake Sato[11], Kazunobu Ouchi[12], Tsutomu Saji[13], Akiyoshi Nariai[14], Yoichi Kaburagi[15], Tetsushi Yoshikawa[16], Kyoko Suzuki[17], Takeo Tanaka[18], Toshiro Nagai[19], Hideo Cho[20], Akihiro Fujino[21], Akihiro Sekine[22], Reiichiro Nakamichi[23], Tatsuhiko Tsunoda[23], Tomisaku Kawasaki[24], Yusuke Nakamura[25] & Akira Hata[1]: [1]Laboratory for Gastrointestinal Diseases, SNP Research Center, RIKEN, Yokohama, Kanagawa, 230-0045, Japan. [2]Department of Hard Tissue Engineering, Graduate School Tokyo Medical and Dental University, 113-8549, Tokyo, Japan. [3]Department of Pediatrics, University of California San Diego, School of Medicine, La Jolla, CA, and Rady Children's Hospital San Diego, CA, USA. [3]Department of Cardiology, Boston Children's Hospital, Boston, MA, USA. [4]Department of Public Health, Jichi Medical School, Minamikawachi, Tochigi, Japan. [4]Saitama Prefectural University, Koshigaya, Saitama, Japan. [5]Department of Preventive Medicine, Shinshu University School of Medicine, Matsumoto, Japan. [6]Department of Molecular Biology, Kawasaki Medical School, Kurashiki, Okayama, Japan. [7]Department of Pediatrics, Fukuoka University School of Medicine, Fukuoka, Fukuoka, Japan. [8]Department of Pediatrics, Tokyo Women's Medical University Yachiyo Medical Center, Yachiyo, Chiba, Japan. [9]Department of Pediatrics, Fuji Heavy Industries LTD. Health Insurance Society General Ohta Hospital, Ohta, Gunma, Japan. [10]Department of Pediatrics, Kawasaki Medical School, Kurashiki, Okayama, Japan. [11]Department of Pediatrics, Toho University School of Medicine, Tokyo, Japan. [12]Department of Pediatrics, Yokohama Minami Kyousai Hospital, Yokohama, Kanagawa, Japan. [15]Department of Pediatrics, National Hospital Organization Yokohama Medical Center, Yokohama, Kanagawa, Japan. [16]Department of Pediatrics, Fujita Health University, Toyoake, Aichi, Japan. [17]Department of Pediatrics, Toyokawa Citizen's Hospital, Toyokawa, Aichi, Japan. [18]Department of Pediatrics, National Hospital Organization Kure Medical Center, Kure, Okayama, Japan. [19]Department of Pediatrics, Dokkyo Medical University, Koshigaya Hospital, Koshigaya, Saitama, Japan. [20]Department of Pediatrics, Kawasaki Municipal Hospital, Kawasaki, Japan. [21]Department of Surgery, Keio University School of Medicine, Tokyo, Japan. [22]Laboratory for Genotyping, SNP Research Center, RIKEN, Yokohama, Kanagawa, Japan. [23]Laboratory for Medical Informatics, SNP Research Center, RIKEN, Yokohama, Kanagawa, Japan. [24]Japan Kawasaki Disease Research Center, Tokyo, Japan. [25]Laboratory for Molecular Medicine, Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, Japan**

Kawasaki disease (KD; OMIM 300530) is an acute, self-limited vasculitis of infants and children. KD is characterized by prolonged fever unresponsive to antibiotics, polymorphous skin rash, erythema of the oral mucosa, lips and tongue, erythema of the palms and soles, bilateral conjunctival injection, and cervical lymphadenopathy. Coronary artery aneurysms develop in 15-25% of untreated patients making KD the

leading cause of acquired heart disease in children in developed countries. Treatment with intravenous immunoglobulin (IVIG) abrogates the inflammation in approximately 80% of patients and reduces the aneurysm rate to less than 5%. Cardiac sequelae of the aneurysms include ischemic heart disease, myocardial infarction, and sudden death. Epidemiological features such as seasonality and clustering of cases suggest an infectious trigger, although no pathogen has been isolated and the etiology remains unknown. Several lines of evidence suggest the importance of genetic factors in disease susceptibility and outcome. First, the incidence of KD is 10 to 20 times higher in Japan than in Western countries[4]. Second, the risk of KD in siblings of affected children is 10 times higher than the general population ($\lambda s=10$) and the incidence of KD in children born to parents with a history of KD is twice as high as the general population. Familial aggregation of the disease has also been observed. Although association studies have identified candidate genes that may influence KD susceptibility, a systematic genetic approach has not been previously performed. Recently, we conducted affected sib pair analysis of KD that demonstrated linkage in several chromosomal regions including chromosome 19. Here we show the results of linkage disequilibrium mapping performed on 19q13.2 which identified a functional SNP in intron 1 of *ITPKC* significantly associated with the risk of KD and the formation of coronary artery aneurysms. We also characterized ITPKC as a novel negative regulator of $Ca^{2+}$/NFAT signaling pathway in T-cells. We identified a functional SNP (itpkc_3) of the inositol 1,4,5-trisphosphate 3-kinase C (*ITPKC*) gene on chromosome 19q13.2 which is significantly associated with susceptibility to KD both in Japanese and U.S. children, as well as with an increased risk of coronary artery lesions. Transfection experiments showed that the C-allele of itpkc_3 reduces splicing efficiency of the ITPKC mRNA. ITPKC acts as a negative regulator of T-cell activation through the $Ca^{2+}$/NFAT signaling pathway and the C-allele may contribute to immune hyper-reactivity in KD. This finding provides new insights into the mechanisms of immune activation in KD and emphasizes the importance of activated T-cells in the pathogenesis of this vasculitis.

### (3) Systemic Lupus Erythematosus

**Tetsuya Oishi[1,2], Aritoshi Iida[3], Shigeru Otsubo[1,2], Yoichiro Kamatani[1], Masayuki Usami[1], Takashi Takei[2], Keiko Uchida[2], Ken Tsuchiya[2], Susumu Saito[3], Yozo Ohnisi[1], Katsushi Tokunaga[4], Kosaku Nitta[2],Yasushi Kawaguchi[5], Naoyuki Kamatani[5], Yuta Kochi[6], Kenichi Shimane[6], Kazuhiko Yamamoto[6], Yusuke Nakamura[1], Wako Yumura[2], and Koichi Matsuda[1*]: [1]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, [4]Department of Human Genetics, Graduate School of Medicine, University of Tokyo, Tokyo 108-8639, Japan. [2]Department of Medicine, Kidney Center, [5]Institute of Rheumatology, Tokyo Women's Medical University, Tokyo 162-8666, Japan. [3]Laboratory for Genotyping, [6]Laboratory for Rheumatic Diseases, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN), Kanagawa 230-0045, Japan.**

SLE (OMIM #152700) is one of the common autoimmune diseases that predominantly afflict women of child-bearing age. The clinical features and serological abnormalities observed in patients with SLE are remarkably diverse, and make clinical assessment and treatment very difficult. Anti-inflammatory drugs and immunosuppressive drugs such as cortisone, azathioprine, methotrexate, and cyclophosphamide have been widely used for treatment of this disease and contribute to significant improvement in prognosis of patients with SLE; although the 4-year survival was estimated to be ~50% in the 1950s (Merrell and Shulman 1955), the 15-year survival rate is at present estimated to be approximately 80%. Despite the improved prognosis in the majority of SLE patients, a significant proportion of the patients still suffer from the disease and/or severe adverse reactions caused by these drugs. For example, synthetic glucocorticoid causes adverse events including hypertension, hyperglycemia, diabetes, cataract, glaucoma, infection, psychosis, osteoporosis, and osteonecrosis. To improve quality of life for the SLE patients, it is one of the critical steps to elucidate the molecular mechanism causing SLE. To identify a gene(s) susceptible to SLE, we performed a case-control association study using genome-wide gene-based single nucleotide polymorphisms (SNPs) in Japanese population. Here we report that an SNP (rs3748079) located in a promoter region of the inositol 1,4,5-triphosphate receptor type 3 (*ITPR3*) gene on chromosome 6p21 was significantly associated with SLE in two independent Japanese case-control samples ($p=0.0000000178$ with odds ratio of 1.88, 95% Confidence Interval (CI) of 1.51-2.35). This particular SNP also revealed associations with rheumatoid arthritis (RA) ($p=0.0084$ with odds ratio of 1.23, 95% CI of 1.05-1.43) and with Graves' disease (GD) ($p=0.00036$ with odds ratio of 1.57, 95% CI of 1.23-2.02). We found the binding of NKX2.5 specific to the

non-susceptible T allele in the region including this SNP. Furthermore, an SNP in *NKX2.5* also revealed an association with SLE ($p=0.0058$ with odds ratio of 1.74, 95% CI of 1.16-2.58). Individuals with risk genotype of both *ITPR3* and *NKX2.5* loci have higher risk for SLE (Odds ratio=5.77). Our data demonstrate that genetic and functional interactions of *ITPR3* and *NKX 2.5* play a crucial role in the pathogenesis of SLE.

We also identified that an SNP, rs3130342, in a 5′ flanking region of the *TNXB* gene revealed a significant association with SLE ($p=0.000000930$, odds ratio (OR) of 3.11 with 95% confidence interval (95%CI) of 1.89-5.28) in Japanese population. This association was replicated by independent 203 cases and 294 controls ($p=0.0440$, OR of 1.52 with 95%CI of 1.01-2.78). Although a copy number variation (CNV) of the *C4* gene adjacent to the *TNXB* gene was reported to be associated with SLE, our analysis on this CNV revealed that the association of CNV of the *C4* gene was weaker than the SNP in the *TNXB* gene and likely to reflect the linkage disequilibrium between *C4* CNV and this particular SNP. Stratified analysis also revealed that the association of SNP rs3130342 with SLE was independent of HLA-DRB1*1501 allele that has been shown to be associated with SLE. Our findings strongly imply that the *TNXB* gene is a candidate gene susceptible to SLE in Japanese population.

## Publications


1. Kato, T., Hayama, S., Yamabuki, T., Ishikawa, N., Miyamoto, M., Ito, T., Tsuchiya, E., Kondo, S., Nakamura, Y. and Daigo, Y. Increased expression of insulin-like growth factor-II messenger RNA-binding protein 1 is associated with tumor progression in patients with lung cancer. Clin Cancer Res. 13: 434-442, 2007.
2. Kanehira, M., Katagiri, T., Shimo, A., Takata, R., Shuin, T., Miki, T., Fujioka, T. and Nakamura, Y. Oncogenic role of MPHOSPH1, a cancer-testis antigen specific to human bladder cancer. Cancer Res. 67: 3276-3285, 2007.
3. Kubo, M., Hata, J., Ninomiya, T., Matsuda, K., Yonemoto, K., Nakano, T., Matsushita, T., Yamazaki, K., Ohnishi, Y., Saito, S., Kitazono, T., Ibayashi, S., Sueishi, K., Iida, M., Nakamura, Y. and Kiyohara, Y.A nonsynonymous SNP in PRKCH increases the risk of cerebral infarction. Nat Genet. 39: 212-217, 2007.
4. Suzuki, C., Takahashi, K., Hayama, S., Ishikawa, N., Kato, T., Ito, T., Tsuchiya, E., Nakamura, Y. and Daigo, Y. Identification of Myc-assocaited protein with JmjC domain as a novel therapeutic target oncogene for lung cancer. Mol Cancer Ther. 6: 542-551, 2007.
5. Yamabuki, T., Takano, A., Hayama, S., Ishikawa, N., Kato, T., Miyamoto, M., Ito, T., Ito, H., Miyagi, Y., Nakayama, H., Fujita, M., Hosokawa, M., Tsuchiya, E., Kohno, N., Kondo, S., Nakamura, Y. and Y. Daigo. Dikkopf-1 as a novel serologic and prognostic biomarker for lung and esophageal carcinomas. Cancer Res. 67: 2517-2525, 2007.
6. Saigusa, K., Imoto, I., Tanikawa, C., Aoyagi, M., Ohno, K., Nakamura, Y. and Inazawa, J. RGC32, a novel p53-inducible gene, is located on centrosomes during mitosis and results in G2/M arrest. Oncogene. 26: 1110-1121, 2007.
7. Ebana, Y., Ozaki, K., Inoue, K., Sato, H., Iida, A., Lwin, H., Saito, S., Mizuno, H., Takahashi, A., Nakamura, T., Miyamoto, Y., Ikegawa, S., Odashiro, K., Nobuyoshi, M., Kamatani, N., Hori, M., Isobe, M., Nakamura, Y. and Tanaka, T. A functional SNP in ITIH3 is associated with susceptibility to myocardial infarction. J Hum Genet. 52: 220-229, 2007.
8. Lin, M.-L., Park, J.-H., Nishidate, T., Nakamura, Y. and Katagiri, T. MELK, maternal embryonic leucine zipper kinase, involved in mammary carcinogenesis through interaction with Bcl-G, a pro-apoptotic member of Bcl-2 family. Breast Cancer Res. 9: R17, 2007.
9. Takata, R., Katagiri, T., Kanehira, M., Shuin, T., Miki, T., Namiki, M., Kohri, K., Tsunoda, T., Fujioka, T. and Nakamura, Y. Validation study of the prediction system for clinical response of M-VAC neoadjuvant chemotherapy. Cancer Sci. 98: 113-117, 2007.
10. Hosokawa, M., Takehara, A., Matsuda, K., Eguchi, H., Ohigashi, H., Ishikawa, O., Shinomura, Y., Imai, K., Nakamura, Y. and Nakagawa, H. Oncogenic role of KIAA0101 interacting with proliferating cell nuclear antigen in pancreatic cancer. Cancer Res. 67: 2568-2576, 2007.
11. Hata, J., Matsuda, K., Ninomiya, T., Yonemoto, K., Matsushita, T., Ohnishi, Y., Saito, S., Kitazono, T., Ibayashi, S., Iida, M., Kiyohara, Y., Nakamura, Y. and Kubo, M. Functional SNP in a Sp1-binding site of AGTRL1 gene is associated with susceptibility to brain infarction. Hum Mol Genet. 16: 630-639, 2007.
12. Osawa, N., Koga, D., Araki, S., Uzu, T.,

Tsunoda, T., Kashiwagi, A., Nakamura, Y. and Maeda, S. Combinational effect of genes for the renin-angiotensin system in conferring susceptibility to diabetic nephropathy. J Hum Genet. 52: 143-151, 2007.
13. Onouchi, Y., Tamari, M., Takahashi, A., Tsunoda, T., Yashiro, M., Nakamura, Yo., Yanagawa, H., Wakui, K., Fukushima, Y., Kawasaki, T., Nakamura, Yu. and Hata, A. A gennomewide linkage analysis of Kawasaki disease:evidence for linkage to chromosome 12. J Hum Genet. 52: 179-190, 2007.
14. Miyamoto, Y., Mabuchi, A., Shi, D., Kubo, T., Takatori, Y., Saito, S., Fujioka, M., Sudo, A., Uchida, A., Yamamoto, S., Ozaki, K., Takigawa, M., Tanaka, T., Nakamura, Y., Jiang, Q. and Ikegawa, S. A functional polymorphism in the 5′ UTR of GDF5 is associated with susceptibility to osteoarthritis. Nat Genet. 39: 529-533, 2007.
15. Kanehira, M., Harada, Y., Takata, R., Shuin, T., Miki, T., Fujioka, T., Nakamura, Y. and Katagiri, T. Involvement of upregulation of DEPDC1 (DEP domain containing 1) in bladder carcinogenesis. Oncogene. 26: 6448-6455, 2007.
16. Hayama, S., Daigo, Y., Yamabuki, T., Hirata, D., Kato, T., Miyamoto, M., Ito, T., Tsuchiya, E., Kondo, S. and Nakamura, Y. Phosphorylation and activation of cell division cycle associated 8 by aurora kinase B plays a significant role in human lung carcinogenesis. Cancer Res. 67: 4113-4122, 2007.
17. Tamura, K., Furihata, M., Tsunoda, T., Ashida, S., Takata, R., Obara, W., Yoshioka, H., Daigo, Y., Nasu, Y., Kumon, H., Konaka, H., Namiki, M., Tozawa, K., Kohri, K., Tanji, N., Yokoyama, M., Shimazui, T., Akaza, H., Mizutani, Y., Miki, T., Fujioka, T., Shuin, T., Nakamura, Y. and Nakagawa, H. Molecular features of hormone-refractory prostate cancer cells by genome-wide gene expression profiles. Cancer Res. 67: 5117-5125, 2007.
18. Bourdon, A., Minai, L., Serre, V., Jais, J.-P., Sarzi, E., Aubert, S., Chretien, D., Lonlay, de P., Paquis-Flucklinger, V., Arakawa, H., Nakamura, Y., Munnich, A. and Rotig, A. Mutation of RRM2B, encoding p53-controlled ribonucleotide reductase (p53R2), causes severe mitochondrial DNA depletion. Nat Genet. 39: 776-780, 2007.
19. Giacomini, K. M., Krauss, R. M., Roden, D. M., Eichelbaum, M., Hayden, M. R. and Nakamura, Y. When good drugs go bad (commentary). Nature. 446: 975-977, 2007.
20. Zembutsu, H., Yanada, M., Hishida, A., Katagiri, T., Tsuruo, T., Sugiura, I., Takeuchi, J., Usui, N., Naoe, T., Nakamura, Y. and Ohno, R. Prediction of risk of disease recurrence by genome-wide cDNA microarray analysis in patients with Philadelphia chromosome-positive acute lymphoblastic leukemia treated with imatinib-combined chemotherapy. Int J Oncol. 31: 313-322, 2007.
21. Fujikawa, M., Katagiri, T., Tugores, A., Nakamura, Y. and Ishikawa, F. ESE-3, an Ets family transcription factor, Is upregulated in cellular senescence. Cancer Sci. 98: 1468-1475, 2007.
22. Hosono, N., Kubo, M., Tsuchiya, Y., Sato, H., Kitamoto, T., Saito, S., Ohnishi, Y. and Nakamura, Y. Multiplex PCR-based real-time Invader assay (mPCR-RTIARETINA): a novel SNP-based detection method for duplicated regions copy number polymorphisms. Hum Mutat. 29: 182-189, 2007.
23. Kato, T., Sato, N., Hayama, S., Yamabuki, T., Ito, T., Miyamoto, M., Kondo, S., Nakamura, Y. and Daigo, Y. Activation of Holliday Junction-Recognizing Protein involved in the chromosomal stability and immortality of cancer cells. Cancer Res. 67: 8544-8553, 2007.
24. Ueda, K., Katagiri, T., Shimada, T., Irie, S., Sato, T., Nakamura, Y. and Daigo, Y. Comparative profiling of serum glycoproteome by sequential purification of glycoproteins and 2-nitrobenzensulfenyl (NBS) stable isotope labeling: a new approach for the novel biomarker discovery for cancer. J Proteome Res. 6: 3475-3483, 2007.
25. Taniwaki, M., Takano, A., Ishikawa, N., Yasui, W., Inai, K., Nishimura, H., Tsuchiya, E., Kohno, N., Nakamura, Y. and Daigo, Y. Activation of KIF4A as a prognostic biomarker and therapeutic target for lung cancer. Clin Cancer Res. 13: 6624-6631, 2007.
26. Suda, T., Tsunoda, T., Daigo, Y., Nakamura, Y. and Tahara, H. Identification of HLA-A 24-restricted epitope-peptides derived from gene products up-regulated in lung and esophageal cancers as novel targets for immunotherapy. Cancer Sci. 98: 1803-1808, 2007.
27. Mano, Y., Takahashi, K., Ishikawa, N., Takano, A., Yasui, W., Inai, K., Nishimura, H., Tsuchiya, E., Nakamura, Y. and Daigo, Y. Growth factor receptor 1 oncogene partner as a noveprognostic biomarker and therapeutic target for lung cancer. Cancer Sci. 98: 1902-1913, 2007.
28. Senju, S., Suemori, H., Zembutsu, H., Uemura, Y., Hirata, S., Fukuma, D., Matsuyoshi, H., Shimomura, M., Haruta, M., Fukushima, S., Matsunaga, Y., Katagiri, T., Nakamura, Y., Furuya, M., Nakatsuji, N. and Nishimura, Y. Genetically manipulated human embryonic stem cell-derived den-

dritic cells with immune regulatory function. Stem Cells. 25: 2720-2729, 2007.
29. Cha, P.-C., Mushiroda, T., Takahashi, A., Saito, S., Shimomura, H., Wanibuchi, Y., Suzuki, T., Kamatani, N. and Nakamura, Y. High-resolution SNP and haplotype maps of the human gamma-glutamyl carboxylase gene (GGCX) and association study between polymorphisms in GGCX with warfarin maintenance-dose requirement of the Japanese population. J Hum Genet. 52: 856-864, 2007.
30. Yamazaki, K., Onouchi, Y., Takazoe, M., Kubo, M., Nakamura, Y. and Hata, A. Association analysis of genetic variants in IL23R, ATG16L1 and 5p13.1 loci with Crohn's disease in Japanese patients. J Hum Genet. 52: 575-583, 2007.
31. Tsukumo, Y., Tomida, A., Kitahara, O., Nakamura, Y., Asada, S., Mori, K. and Tsuruo, T. Nucleobindin 1 controls the unfolded protein response by inhibiting ATF6 activation. J Biol Chem. 282: 29264-29272, 2007.
32. Takehara, A., Hosokawa, M., Eguchi, H., Ohigashi, H., Ishikawa, O., Nakamura, Y. and Nakagawa, H. GABA stimulates pancreatic cancer growth through over-expressing GABAA receptor pai subunit (GABRP). Cancer Res. 67: 9704-9712, 2007.
33. Kunizaki, M., Hamamoto, R., Silva, F.P., Yamaguchi, K., Nagayasu, T., Shibuya, M., Nakamura, Y. and Furukawa, Y. The lysine 831 of VEGFR1 a novel target of methylation by SMYD3. Cancer Res. 67: 10759-10765, 2007.
34. Onouchi, Y., Gunji, T., Burns, J.C., Shimizu, C., Newburger, J.W., Yashiro, M., Nakamura, Yo., Yanagawa, H., Wakui, K., Fukushima, Y., Kishi, F., Hamamoto, K., Terai, M., Sato, Y., Ouchi, K., Saji, T., Nariai, A., Kaburagi, Y., Yoshikawa, T., Suzuki, K., Tanaka, T., Nagai, T., Cho, H., Fujino, A., Sekine, A., Nakamichi, R., Tsunoda, T., Kawasaki, T., Nakamura, Yu. and Hata, A. A functional polymorphism in ITPKC is associated with Kawasaki disease susceptibility and formation of coronary artery aneurysms. Nat Genet. 40: 35-42, 2008.
35. Kudo, S., Konda, R., Obara, W., Kudo, D., Tani, K., Nakamura, Y. and Fujioka, T. Inhibition of tumor growth through suppression of angiogenesis by brain-specific angiogenesis inhibitor 1 gene transfer in murine renal cell carcinoma. Oncol Rep. 18: 785-791, 2007.
36. Silva, F.P., Hamamoto, R., Kunizaki, M., Tsuge, M., Nakamura, Y. and Furukawa, Y. Enhanced methyltransferase activity of SMYD3 by the cleavage of its N-terminal region in human cancer cells. Oncogene, Enhanced methyltransferase activity of SMYD3 by the cleavage of its N-terminal region in human cancer cells. Oncogene. Nov 12; [Epub ahead of print], 2007.
37. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 449: 851-861, 2007.
38. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., Schaffner, S.F., Lander, E.S. and The International HapMap Consortium: Genome-wide detection and characterization of positive selection in human populations. Nature. 449: 913-918, 2007.
39. Ishikawa, N., Takano, A., Yasui, W., Inai, K., Nishimura, H., Ito, H., Miyagi, Y., Nakayama, H., Fujita, M., Hosokawa, M., Tsuchiya, E., Kohno, N., Nakamura, Y. and Daigo, Y. Cancer-testis antigen, LY6K is a serologic biomarker and a therapeutic target for lung and esophageal carcinomas. Cancer Res. 67: 11601-11611, 2007.
40. Kato, M., Miya, F., Kanemura, Y., Tanaka, T., Nakamura, Y. and Tsunoda, T. Recombination rates of genes expressed in human tissues. Hum Mol Genet. Nov 13; [Epub ahead of print], 2007.
41. Kidokoro, T., Tanikawa, C., Furukawa, Y., Katagiri, T., Nakamura, Y. and Matsuda, K. CDC20, a potential cancer therapeutic target, is negatively regulated by p53. Oncogene. Sep 17; [Epub ahead of print], 2007.
42. Brunet, J., Pfaff, A.W., Abidi, A., Unoki, M., Nakamura, Y., Guinard, M., Klein, J.-P., Candolfi, E. and Mousli, M. Toxoplasma gondii exploits UHRF1 and induces host cell cycle arrest at G2 to enable its proliferation. Cell Microbiol. Dec 6; [Epub ahead of print], 2007.
43. Shimo, A., Tanikawa, C., Nishidate, T., Matsuda, K., Lin, M.-L., Park, J.-H., Ohta, T., Hirata, K., Fukuda, M., Nakamura, Y. and Katagiri, T. Involvement of KIF2C/MCAK overexpression in mammary carcinogenesis. Cancer Sci. 99: 62-70, 2008.
44. Leung, A.A.C.C., Wong, V.C.L., Yang, L.C., Chan, P.L., Daigo, Y., Nakamura, Y., Qi, R. Z., Miller, L., Liu, E.T.-K., Wang, L.D.J.-L., S. Law, Tsao, W. and Lung, M.L. Frequent decreased expression of candidate tumor suppressor gene, DEC1, and its anchorage-independent growth properties and impact on global gene expression in esophageal carcinoma. Int J Cancer. 122: 587-594, 2008.
45. Uemura, M., Tamura, K., Chung, S., Honma, S., Okuyama, A., Nakamura, Y. and Nakagawa, H. A novel 5-steroid reductase (SRD5

A3, type-3) is overexpressed in hormone-refractory prostate cancer. Cancer Sci. 99: 81-86, 2008.
46. Kamatani, Y., Matsuda, K., Ohishi, T., Ohtsubo, S., Yamazaki, K., Iida, A., Hosono, N., Kubo, M., Yumura, W., Nitta, K., Katagiri, T., Kawaguchi, Y., Kamatani, N. and Nakamura, Y. Identification of a significant association of an SNP in TNXB with SLE in Japanese population. J Hum Genet. 53: 64-73, 2008.
47. Obama, K., Satoh, S., Hamamoto, R., Sakai, Y., Nakamura, Y. and Furukawa, Y. Enhanced expression of RAD51AP1 is involved in the growth of intrahepatic cholangiocarcinoma cells. Clin Cancer Res. in press, 2008.
48. Fukukawa, C., Hanaoka, H., Nagayama, S., Tsunoda, T., Toguchida, J., Endo, K., Nakamura, Y. and Katagiri, T. Radioimmunotherapy of human synovial sarcoma using a monoclonal antibody against FZD10. Cancer Sci. in press, 2008.
49. Kato, N., Miyata, T., Tabara, Y., Katsuya, T., Yanai, K., Hanada, H., Kamide, K., Nakura, J., Kohara, K., Takeuchi, F., Mano, H., Yasunami, M., Kimura, A., Kita, Y., Ueshima, H., Nakayama, T., Soma, M., Hata, A., Fujioka, A., Kawano, Y., Nakao, K., Sekine, A., Yoshida, T., Nakamura, Y., Saruta, T., Ogihara, T., Sugano, S., Miki, T. and Tomoike, H. High-Density Association Study and Nomination of Susceptibility Genes for Hypertension in the Japanese National Project. Hum Mol Genet. in press, 2008.
50. Oishi, T., Iida, A., Otsubo, S., Kamatani, Y., Usami, M., Takei, T., Uchida, K., Tsuchiya, K., Saito, S., Ohnishi, Y., Tokunaga, K., Nitta, K., Kawaguchi, Y., Kamatani, N., Kochi, Y., Shimane, K., Yamamoto, K., Nakamura, Y., Yumura, W. and Matsuda, K. A functional SNP in the NKX2.5-binding site of ITPR3 promoter is associated with susceptibility to Systemic Lupus Erythematosus in Japanese population. J Hum Genet. 53: 151-162, 2008.
51. Daigo, Y. and Nakamura, Y. From cancer genomics to the thoracic oncology: New biomarker and therapeutic target discovery for lung and esophageal carcinomas. (Review Article) Gen Thorac and Cardiovasc Surg. 56: 43-53, 2008.

*Human Genome Center*

# Laboratory of Functional Genomics
# ゲノム機能解析分野


| | | | | | |
|---|---|---|---|---|---|
| Visiting Professor | Gregory Mark Lathrop, Ph.D. | | 客員教授 | 理学博士 | グレゴリー・マーク・ラスロップ |
| Associate Professor | Ryo Yamada, M.D., Ph.D. | | 准教授 | 医学博士 | 山　田　　亮 |



*Genetic heterogeneity of human beings is one of the most important targets of post-genomic research. Genome-wide association studies are being actively carried out using the genetic polymorphism markers to identify disease-related loci. We focus on the development of new methods to interpret the heterogeneity and to map the disease-associated loci and collaborate with research groups for data-mining of their genetic epidemiology studies.*


## 1. The development of new methods to map disease-associated loci with genetic polymorphisms.

**Ryo Yamada**

Genome-wide association (GWA) studies are resulting in many useful findings. The scale of such studies is increasing along with rapid progress in genotyping technology. This increase in scale necessarily increases the degree of dependence among individual tests in GWA studies. The inter-test dependence is problematic because almost all the conventional statistical methods assume independence among multiple tests. Besides the multiple sources of inter-test dependency, the variable inflation of test statistics due to biased sampling from structured population is one of the unavoidable consequences of enlarged sample size. These problems that complicate the interpretation of data of GWA studies are mutually related and there is no straight-forward solution of them all together. We decompose the difficulty into parts, i.e., the problem of linkage disequilibrium (LD), population structure, multiple genetic models, study design and characterize their problem and propose solution of the individual problems at the beginning and also attempt to improve the interpretation of data of GWA studies as a whole.

### a. Multiple testing correction

One of the major sources of inter-test dependence is allelic association due to LD and population structure. This problem has been well aware of since SNP-based LD mapping became popular. There are some other origins of inter-test dependence that are less noticed but similarly troublesome as allelic association; the inter-test dependence due to application of multiple genetic models to individual markers and due to study designs, such as the usage of common controls and subjects with comorbidity for multi-phenotype studies, which is currently common particularly in the gigantic study design. This year we developed a method to estimate the effective number of independent tests from GWA studies in the structured population.

### b. Test statistics correction for data of structure population

Because the genetic epidemiology studies on complex genetic traits target relatively weak fac-

tors, which means sample size of them should be more than thousands and subsequently makes idealistic random sampling from homogeneous population impossible. The test statistics of the studies in the heterogeneous population, in other words, structured population, tends to give false positive results. One of the methods to correct the increase in the false positives is genomic control method for chi-square distribution. We modified the genomic control method so that it could correct the Fisher's exact test statistics in 2007.

### c. Characterization of exact trend test for SNP case-control association test data

The trend test is the test of the choice for 2×3 contingency table test of SNP data. The definition of the two-sided exact trend test is controversial. We investigated the suitability of multiple ways of the two-sided exact trend test for 2 ×3 SNP data tables this year.

## 2. The development of new methods to interpret the genetic heterogeneity.

**Ryo Yamada**

As a compound in nature, the DNA sequence is under pressure to maximize the heterogeneity of the sequence. Under the most random condition, all bases of the sequence would be polymorphic, and all bases and all sets of bases are mutually independent. At the other extreme, under the least random condition, all DNA molecules would be clones. In living organisms, the number of polymorphic sites in the DNA sequence is limited due to the requirements for reproduction and as a result of selection and genetic drift, against which opposite forces act to increase heterogeneity (e.g., mutation and recombination). A major research target following the completion of the genome sequence is the investigation of intra-species variations, among which diallelic single nucleotide polymorphisms are the most common.

### a. Quantitation of linkage disequilibrium of multiple markers

Genetic variations within a population give rise to LD, and the use of the genetic history of the population and LD mapping is a very promising method for identifying genetic backgrounds of various phenotypes. LD is a measure of inter-marker dependence. Although the inter-marker dependence exist among any set of markers, only the pair-wise inter-marker dependence is utilized for quantitation of the genetic heterogeneity and for genetic epidemiology studies usually. We develop a new method to quantify the heterogeneity and complexity of population of DNA sequence with SNPs so that various researches based on genetic heterogeneity will benefit in 2007.

### b. Geometric expression of haplotype populations

Haplotypes are consisted of alleles of multiple markers. We attempted to deal the haplotype data from combination theory standpoint and investigated the utility of polyhedral handling of the combinatorial aspects of haplotypes.

## 3. Collaboration with genetic epidemiology research groups.

**Gregory Mark Lathrop and Ryo Yamada**

Besides the development of new methods to analyze genetic polymorphism data in the context of population genetics and genetic statistics, we collaborate with multiple research groups in and out of the IMS-UT, including Kyoto University, Kyoto, The University of Tokyo Hospital, Tokyo, Laboratory for Rheumatic Diseases, SRC, RIKEN, Yokohama, National Hospital Organization Sagamihara National Hospital, Sagamihara, and The Centre National de Génotypage, Evry, France, for the interpretation of genetic epidemiology data with the conventional statistical methods.

## 4. Public distribution of population genetics and genetic association study tools.

**Ryo Yamada**

Because the designs of genetic epidemiology studies have been changing, the analysis tools have to be updated all the time. The number of genetic epidemiology study groups is much more than the groups on genetic statistics in the world and also in Japan. We opened the web site that distributes basic tool of linkage disequilibrium mapping for public use. This distribution is supported by the grant from Japan Society for the Promotion of Science on the permutation test.
Web-site URL: http://func-gen.hgc.jp/

## Publications


1. Yamamoto, K. and Yamada, R. Lessons from a Genomewide Association Study of Rheumatoid Arthritis. N. Engl. J. Med. 357: 1250-1251, 2007
2. Yamada, R. and Yamamoto, K. Mechanisms of disease: genetics of rheumatoid arthritis--ethnic differences in disease-associated genes. Nat. Clin. Pract. Rheumatol. 3: 644-50, 2007
5. Suzuki, A., Yamada, R. and Yamamoto, K. Citrullination by peptidylarginine deiminase in rheumatoid arthritis. Ann. N.Y. Acad. Sci. 1108: 323-39, 2007
6. Ryo Yamada, and F. Matsuda. A novel method to express SNP-based genetic heterogeneity, $\Psi$, and its use to measure linkage disequilibrium for multiple SNPs, Dg, and to estimate absolute maximum of haplotype frequency. Genetic Epidemiol. 31: 709-726, 2007

## *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 |
| Associate Professor | Kengo Kinoshita, Ph.D. | 准教授　理学博士 | 木　下　賢　吾 |


*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 on 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. Transcription factor binding site co-occurrence in firmicutes

**Nicolas Sierro and Kenta Nakai**

Preliminary investigation of binding site co-occurrence in firmicutes highlighted several specific combinations of transcription factors appearing with a higher-than-expected frequency in some genomes. Such combinations could indicate the presence of different regulation mechanisms for specific pathways and in specific strains. For example, two LexA sites separated by 16 or 17 nucleotides are found about 150 nucleotides upstream of *lexA* and a gene similar to *uvrX* in *Bacillus anthracis* and *clausii, Listeria monocytogenes* and *Staphylococcus aureus, epidermidis*, *haemolyticus* and *saprophyticus* strains. Both of these genes are repressed by the SOS DNA-repair genes repressor LexA. However, with the notable exception of the *Staphylococcus* strains for which a similar pattern with a distance of 31 nucleotides is found in front of *recA*, no other gene of the SOS DNA-repair regulon is preceded by such combinations. This specific co-occurrence pattern could be related to the two-step SOS DNA-repair response mechanism, where the presence of multiple LexA sites upstream of some of the genes involved in the first response may be necessary to ensure a tight repression of potentially harmful genes despite the weaker binding site.

### 2. Promoter structure modeling for expression pattern prediction

**Alexis Vandenbon and Kenta Nakai**

Initiation of transcription in eukaryotes is regulated by transcription factors binding *cis*-regulatory elements in the regulatory regions of genes. We can therefore assume that regulatory regions containing similar sets of *cis*-regulatory motifs will drive similar expression patterns at the transcriptional level. We have developed a Markov chain-based promoter structure model that includes positional preference, order, and orientation of motifs in upstream regulatory regions. The model is trained using promoter sequences of a set of co-regulated genes, and is subsequently used to predict genes having similar expression profiles as the input gene set. We applied our model on a dataset of genes expressed in pharyngeal muscle cells of *Caenorhabditis elegans*, and on a dataset of muscle expressed genes of *Ciona intestinalis*. Both compu-

tational and experimental verifications indicated that this model is capable of predicting candidate promoters driving similar expression patterns as the input-regulatory sequences. We are further developing our promoter structure model, and we are also working on a number of projects to find common features in sets of co-regulated genes. We believe that our approaches can be useful for finding promising candidate genes for wet-lab experiments and for increasing our understanding of transcriptional regulation.

## 3. DBTSS: database of transcription start sites, progress report 2008

**Hiroyuki Wakaguri[1], Riu Yamashita, Yutaka Suzuki[1], Sumio Sugano[1], and Kenta Nakai: [1]Graduate School of Frontier Sciences, U. Tokyo**

One of the major topics in the post-sequence era is the analysis of transcription regulation network. In order to study this problem, we have constructed a database, DataBase of Transcription Starts Site (DBTSS), which included a number of 5'-end sequence (ESTs) produced by the oligo-capping method. We have recently released DBTSS version 6, extended with three major additional features. The first new feature is the new data resource by SOLEXA. Now DBTSS has not only 1,540,411 mapped sequences from oligo-capped clones, but also 21,981,890 SOLEXA sequences from MCF7 and HEK193 cells. The second feature is a viewer of evolutional conservation of the promoter regions. A user can compare the sequence around the promoter regions from three species among human, mouse, rat, chimp, and macaque. The third feature is a tool for the transcriptome analysis of promoter region with expression and predicted transcription factor binding sites. This tool searches for putative transcription factor binding sites commonly occurring in the promoters, which show similar behaviors on various drug perturbations. DBTSS can be accessed at http://dbtss.hgc.jp.

## 4. Predicting transcriptional activities of human promoters from putative transcription factor binding sites by using support vector machines

**Hirokazu Chiba, Riu Yamashita, Kenta Nakai**

Bindings of specific transcription factors (TFs) to promoter sequences play a central role in transcription regulation. However, the relationship between the combination of TFs and transcription activities remains to be deciphered. Toward the understanding the relationship, we analyzed human promoter sequences and their transcriptional activities in specific tissues, which were defined by large collection of full-length cDNAs. We used support vector machine to predict the transcriptional activities from putative TF binding sites (TFBSs) in the promoter sequences. Among 38 tissues tested, best prediction performances were obtained in the cases of brain and testis. Furthermore, we performed the same scheme of prediction by dividing the promoter sequences into two parts at -200 of transcriptional start sites (TSSs). The prediction performances were improved, suggesting that the TFBSs function differently according to the position, and the boundary is -200. We also performed cross-species comparison of promoter sequences between human and mouse, and examined the average identities at various positions from the TSSs. The result supports the functional boundary at -200 of TSSs. These results will promote the understanding of the promoter architecture of mammals.

## 5. Retrotransposition as a source of new promoters

**Kohji Okamura and Kenta Nakai**

The fact that promoters are essential for the function of all genes presents the basis of the general idea that retrotranspositions give rise to processed pseudogenes. However, recent studies have demonstrated that some retrotransposed genes are transcriptionally active. Because promoters are not thought to be retrotransposed along with exonic sequences, these transcriptionally active genes must have acquired a functional promoter by mechanisms that are yet to be determined. Hence, comparison between a retrotransposed gene and its source gene appears to provide a unique opportunity to investigate the promoter creation for a new gene. We identified 29 gene pairs in the human genome, consisting of a functional retrotransposed gene and its parental gene, and compared their respective promoters. In more than half of these cases, we unexpectedly found that a large part of the core promoter had been transcribed, reverse-transcribed, and then integrated to be operative at the transposed locus. This observation can be ascribed to the recent discovery that transcription start sites tend to be interspersed rather than situated at one specific site. This propensity could confer retrotransposability to promoters per se. Accordingly, the retrotransposability can explain the genesis of some alternative promoters.

## 6. Investigation of motif finding performances

**Keishin Nishida and Kenta Nakai**

Computational prediction of nucleotide binding specificity for transcription factors remains a fundamental and largely unsolved problem. One of the prediction methods is pattern discovery in unaligned DNA sequences, called "motif finding". Despite many studies about motif finding, this problem is far from being resolved. To provide an appropriate motif finding overview for experiment design, we tried larger scale analysis of wide input conditions. We focused on the Gibbs Motif Sampler, one of the most famous motif finding programs. Background sequences are generated randomly with any GC content. Motif information is downloaded from the JASPAR database to plant experimentally valid motif into the background sequences. Our results show an expected tendency; longer sequences are a more difficult problem, and more mismatches further increase the difficulty. We find that the motif finding performance is worse for larger input sequences than for smaller ones. Indicating that the default program parameters are not optimal for larges. Based on this result, we will investigate suitable Gibbs Motif Sampler parameters for handling such large size sequence datasets. In addition we will investigate other motif models and other motif finding programs, and produce similar guidelines.

## 7. Analysis of Nucleosomal DNA sequences

**Yoshiaki Tanaka, Riu Yamashita and Kenta Nakai**

Nucleosomes limit accessibility of some regulatory factor binding sites, and thus play an important role in transcription and replication. It has been reported that the nucleosome occupancy rate in the regions upstream of transcription start sites (TSSs) is lower than that in other regions. It is also known that the binding of nucleosomes on DNA is sequence dependent. Motifs recognized by nucleosomes are categorized into two groups, periodic sequence motifs (ex. 10bp AA/TT repeat) and consecutive sequence motifs (ex. poly(A/T)). Recently some researchers succeeded in the computational prediction of nucleosome positions in the *S. cerevisiae* genome using some of these motifs. However, their methods are not as efficient in higher eukaryotic genomes. We therefore compared nucleosome and linker DNA sequences from the genomes such as *H. sapiens, C. elegans* and *S. cerevisiae*. In this analysis, different tendencies are observed for each organism. For example, 10bp AA/TT periodic motifs are observed in *C. elegans* and *S. cerevisiae*, but not in *H. sapiens*. These tendencies will be helpful for developing a new prediction approach of nucleosome positioning. Now we are also investigating other important tendencies for nucleosome positioning in higher eukaryotic genomes.

## 8. Computational analysis of trans-splicing in *C. intestinalis* gene expression

**Shuang Li, Riu Yamashita, Nicolas Sierro and Kenta Nakai**

Trans-splicing, in which the 5′-ends of some pre-mRNAs are cut and replaced by the 5′-terminal sequence of a specific donor mRNA called spliced-leader (SL), has been observed in six phyla. In chordates, *Ciona intestinalis* is the only one reported as performing trans-splicing systematically. We have studied the trans-splicing of *C. intestinalis* with 5′-EST data. This year, we compared the two populations (trans-splicing positive or negative) of genes, aiming at finding the biological meaning of trans-splicing. We discovered that the expression ratio between the two gene groups varies with tissues and developmental stages, ranging from 1:3.1 at the 'juvenile' stage to 1:1.1 in 'blood'. We also observed for instance that ribosome related genes are more likely to fall into the non-trans-spliced group while mitochondria related genes show a preference for the trans-spliced group. Our analysis indicates that in *C. intestinalis*, although there may not exist strong fundamental requirements for genes to generate trans-splicing pre-mRNAs, the different populations of genes are likely to be spatially and temporally regulated differently.

## 9. Computational Description of Gene Networks by Direct Regulatory Interactions in Ascidian Early Development

**Xuyang Yuan, Nicolas Sierro, Kohji Okamura, and Kenta Nakai**

The ascidian *Ciona* is a good model system to elucidate gene regulatory networks in chordate development. Accordingly, comprehensive *in situ* hybridization assays have identified a number of regulatory genes with localized expression pattern. Subsequent knockdown assays have illuminated thousands of combinations of gene expression profiles in the early embryo, depicting a blueprint for its early development. However, the blueprint cannot demonstrate di-

rect molecular mechanisms occurring in each cell because an interaction between two genes was detected by whether each transcription takes place or not. Thus, we are computationally constructing gene networks consisting of only direct interactions. For *trans*-acting factors whose binding sites are unknown, we are predicting them by motif finding programs and examination of orthologs to complete the whole network. Our result will help to understand the precise molecular mechanisms during the development and will provide suggestion for further experimental analyses.

## 10. Docking of protein molecular surfaces with evolutionary trace analysis

**Eiji Kanamori[2,3], Yoichi Murakami[4], Yuko Tsuchiya, Daron M Standley[4], Haruki Nakamura[4], and Kengo Kinoshita: [2]Japan Biological Information Research Center, [3]Hitachi Software Engineering Co., Ltd., [4]Protein Research Institute, Osaka University**

Protein-protein interaction is the first step to construct interaction network of proteins, which is a key to understand the complex biological functions. To identity the reliable interactions, structural information of proteins are useful, so the method to predict the complex structure of proteins is require. Many methods have been developed by several groups, but the prediction accuracy is not so high at the moment. Therefore, we tried to develop a new method to predict the complex structure of proteins using the different view of proteins, that is, molecular surface with considering the physicochemical and evolutional information. With this method, we participated in blind prediction contest, critical assessment of prediction of interactions (CAPRI) and have achieved the good results.

## 11. COXPRESdb: a database of coexpressed gene networks in mammals

**Takeshi Obayashi, Shinpei Hayashi[5], Masayuki Shibaoka1, Motoshi Saeki[5], Hiroyuki Ohta[5,6], Kengo Kinoshita: [5]Graduate School of Information Science and Engineering, Tokyo Institute of Technology, [6]Center of Biological Resources and Informatics, Tokyo Institute of Technology**

The amount of publicly available gene expression data is the most abundant in mouse and human, which is tenth or twentieth larger than that for Arabidopsis. However there is no coexpressed gene database as like ATTED-II for Arabidopsis, although the information is valuable to predict gene function. We are thus constructing new database named COXPRESdb (coexpression database) (http://coxpresdb.hgc.jp) for coexpressed genes in mouse and human from such publicly available gene expression data. The information of gene coexpression is calculated from thousands of oligonucleotide microarray (GeneChip) data and then represented as gene lists and gene networks. When the networks became too big for the static picture on the web in GO networks or in tissue networks, we used Google Maps API to visualize them interactively. This information of gene coexpression will widely promote experimental researches for mouse and human.

## 13. Identification of transient hub proteins and the possible structural basis for their multiple interactions

**Miho Higurashi, Takashi Ishida and Kengo Kinoshita**

Proteins that can interact with multiple partners play central roles in the important biological processes, such as signal transduction. Although it is well known that stable complexes and transient complexes have different structural features, the hub proteins defined by previous study were identified as proteins with multiple partners, regardless of whether the interactions were transient or stable. As a result, so-called hub proteins can be the components of stable supramolecules as opposed to the normal concepts of hub proteins. In this study, we have developed a method to identify transient hub proteins using PDB, and then perform statistical analyses of the structural features of these proteins. As a result, we found that the main difference between sowciable and non-sociable proteins is not the abundance of disordered regions, in contrast to the previous studies, but rather the structural flexibility of the entire protein, resulting from less number of hydrogen bonds between core residues.

## 14. PrDOS: prediction of disordered protein regions from amino acid sequence

**Takashi Ishida and Kengo Kinoshita**

Identification of disordered regions in proteins is important for the functional annotation of proteins and for high-throughput structural determination. However, the cost of experimental determination of disordered regions is expensive. Thus, we developed a computational method to predict disordered protein regions from their amino acid sequences and imple-

mented web-interface of this prediction method. Our system is composed of two predictors, that is, a predictor based on the local amino acid sequence, and one based on template proteins. The performance of the method was evaluated in the blind benchmark by the structural biology community. The method achieved high accuracy (>90% with the sensitivity of 0.56), especially for short disordered regions.

### 15. eF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape

**Kengo Kinoshita, Yoichi Murakami[4] and Haruki Nakamura[4]**

Molecular function of proteins are determined by their three dimensional structures, thus the similarity of protein structure can give some clues to infer their functions. In many cases, the molecular functions are begun with the molecular interaction with small molecules (ligands). Therefore, to find the putative ligands is the first step to identify the molecular function of proteins. For the purpose, we have developed a web server to identify a putative ligand upon the structure of proteins. The web server, eF-seek, accepts a coordinate file with PDB format file and returns complex structures predicted. to search for the similar ligand binding sites for the uploaded coordinate file with PDB format. The representative binding sites in eF-site database are search by our own algorithm based on the clique search algorithm.

### 16. Structural diversity of ligand binding sites in proteins

**Megumi Okamoto, Matsuyuki Shirota, and Kengo Kinoshita**

Molecule recognition is the first step in realizing the function of proteins. Therefore, the understanding of the molecular recognition is an important step to reveal the structure-function relation in proteins. To observe the structure-function relationship of known complexes, we carried out the classification of AMP, GMP, CMP, UMP binding sites in PDB. We analyzed variety of binding sites in three levels, which are secondary structure elements, spatial arrangement of atoms and shape of three-residue fragments. As a result, we found that binding sites, which are various in the view of secondary structure and atomic configuration, have common fragment combinations. And same fragment combination constitutes structurally different binding sites by different spatial arrangements.

### 17. Molecular dynamics simulation of cholesterol-induced change in a membrane environment

**Naoya Fujita, Takashi Ishida and Kengo Kinoshita**

Simulation in molecular level of lipid raft, where many biological functions are realized, is important to understand membrane heterogeneity, function of proteins and effects of raft structure to proteins' function. To reveal effects of cholesterol, which is one of main components in the raft domain, we compared a cholesterol-mixed raft-like membrane and a pure lipid, dipalmitoylphosphatidylcholine (DPPC), membrane. By analyzing the simulation trajectories, a significant variation of membrane width was found on the pure DPPC membrane but not on the cholesterol-mixed membrane. Such a difference on the membrane environment also changed stability and mobility on an embedded protein, alamethicin.

### 18. Domain size effect on amino acid composition

**Matsuyuki Shirota, Takashi Ishida, and Kengo Kinoshita**

The size effect of proteins on amino acid composition, that is whether the frequencies of residues change systematically with increasing protein size, has been tested several times under the assumption that the frequencies of hydrophobic residues would increase and those of hydrophilic residues would decrease if protein size increased. However, such systematic change in the frequency of hydrophobic or hydrophilic residues has not yet been detected. Here, taking advantage of recent large database of protein structures, we examined the change in occurrence of each amino acid with increasing protein size, and identified definite correlation between protein size and occurrence of several amino acids. Contrary to the generally accepted expectation that the incidence of hydrophilic amino acids would decrease, only two charged amino acids-lysine and glutamic acid-among all hydrophilic residues had decreased occurrence with increasing protein size. Since these two amino acids are the most rarely found in the protein core, this decrease of lysine and glutamic acid can be considered to be induced by the increase of protein size due to the decrease of relative surface area.

## Publications


Vandenbon, A., Miyamoto, Y., Takimoto, N., Kusakabe, T., and Nakai, K. Markov chain-based promoter structure modeling for tissue-specific expression pattern prediction. *DNA Res*., in press.

Genome Information Integration Project and H-invitational 2 Consortium, The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucl. Acids Res*., in press.

Sierro, N., Makita, Y., de Hoon, M., and Nakai, K. DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucl. Acids Res*., in press.

Wakaguri, H., Yamashita, R., Suzuki, Y., Sugano, S., and Nakai, K. DBTSS: DataBase of Transcription Start Sites, progress report 2008. *Nucl. Acids Res*., in press.

Ikura, T., Kinoshita, K., and Ito, N.A cavity with an appropriate size is the basis of the PPIase activity. *PEDS*, in press.

Obayashi, T., Hayashi, S., Shibaoka, M., Saek, M., Ohta, H., and Kinoshita, K. COXPRESdb: a database of coexpressed gene networks in mammals. *Nucl. Acids Res*, in press.

Higurashi, M., Ishida, T., and Kinoshita, K. Identification of transient hub proteins and the possible structural basis for their multiple interactions. *Protein Sci*, 17: 72-78, 2008.

Kanamori, E., Murakami, Y., Tsuchiya, Y., Standley, D.M., Nakamura, H., and Kinoshita, K. Docking of protein molecular surfaces with evolutionary trace analysis. *Proteins* 69: 832-838, 2007.

Okumura, T., Makiguchi, H., Makita, Y., Yamashita, R., and Nakai, K. Melina II: a web tool for comparisons among several predictive algorithms to find potential motifs from promoter regions. *Nucl. Acids Res*. 35: W227-W 231, 2007.

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J., and Nakai, K. WoLF PSORT: protein localization predictor. *Nucl. Acids Res*. 35: W585-W587, 2007.

Tsuritani, K., Irie, T., Yamashita, R., Wakaguri, H., Kanai, A., Mizushima-Sugano, J., Sugano, S., Nakai, K., and Suzuki, Y. Distinct class of putative "non-conserved" promoters in humans: comparative studies of alternative promoters of human and mouse genes. *Genome Res*. 17(7): 1005-1014, 2007.

Sakakibara, Y., Irie, T., Suzuki, Y., Yamashita, R., Wakaguri, H., Kanai, A., Chiba, J., Takagi, T., Mizushima-Sugano, J., Hashimoto, S., Nakai, K., and Sugano, S. Intrinsic promoter activities of primary DNA sequences in the human genome. *DNA Res*. 14(2): 71-77, 2007.

Nakai, K. and Horton, P. Chapter 29: Computational prediction of subcellular localization (in M. van der Giezen ed., *Protein Targeting Protocols* (2nd ed.), *Methods in Molecular Biology*, vol. 390). pp. 429-466, Humana Press, Totowa, 2007, (ISBN 13 978-1-58829-702-0).

Obayashi, T., Kinoshita, K., Nakai, K., Shibaoka, M., Hayashi, S., Saeki, M., Shibata, D., Saito, K., and Ohta, H. ATTED-II: a database of co-expressed genes and *cis* elements for identifying co-regulated gene groups in *Arabidopsis*. *Nucl. Acids Res*. 35: D863-D869, 2007.

Koike, R., Kinoshita, K., and Kidera, A. Probabilistic alignment detects remote homology in a pair of protein sequences without homologous sequence information. *Proteins* 66: 655-63, 2007.

Kinoshita, K., Murakami, Y., and Nakamura, H. eF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape. *Nucleic Acids Res*. 35: W398-402, 2007.

Ishida, T. and Kinoshita, K. PrDOS: prediction of disordered protein regions from amino acid sequence, *Nucleic Acids Res*. 35: W460-464, 2007.

Hosaka, Y., Iwata, M., Kamiya, N., Yamada, M., Kinoshita, K., Fukunishi, Y., Tsujimae, K., Hibino, H., Aizawa, Y., Inanobe, A., Nakamura, H., and Kurachi, Y. Mutational analysis of block and facilitation of HERG current by a class III anti-crrhythmic agent, nifekalant. *Channels* 1: 198-208, 2007.

蒔田由布子，中井謙太．第２章10.転写因子とシスエレメントの相互作用解析に役立つデータベースとウェブツール．磯辺・中山・伊藤編，分子間相互作用解析ハンドブック(実験ハンドブックシリーズ)．羊土社．pp. 210–217, 2007.

鈴木穣，山下理宇，中井謙太，菅野純夫．転写制御領域(プロモーター領域)の同定と解析～DBTSS/TRANSFAC～．中村・磯合・石川・平川・坊農編，バイオデータベースとウェブツールの手とり足とり活用法　改訂第２版．羊土社．pp. 65–75, 2007.

ポール・ホートン，中井謙太．PSORT(ピーソート)．中村・磯合・石川・平川・坊農編，バイオデータベースとウェブツールの手とり足とり活用法　改訂第２版．羊土社．pp. 108–114, 2007.

木下賢吾．タンパク質間相互作用に関するデータベースの構築．生体の科学. 58：334–337, 2007.

*Human Genome Center*

# Department of Public Policy
## 公共政策研究分野

Associate Professor Kaori Muto, Ph.D.

准教授 保健学博士 武 藤 香 織

*Department of Public Policy is a new and the first social science section at the IMSUT. We work for three major missions; public policy studies on translational research, its application to healthcare and its impact on social security; practical advices and survey for research projects to build public trust; and "minority-centered" scientific communication. We have conducted a comparative political study on genetic testing business in East Asia and supported for "BioBank Japan" project from ethical, legal and social standpoints.*

### 1. A comparative political study on genetic testing business in East Asia

To examine broader social and cultural agendas on industrialization of genetic tests and suggest policy implications in East Asia, we've been conducting this research. In the first phase, we have surveyed political and legal positions regarding genetic testing business, especially genetic tests related to lifestyle and multi-factorial common diseases directly to the public or via clinics in Japan, China, Taiwan and South Korea. We've interviewed main players in this area; the relevant authorities, bioindustries, physicians, academics and patients support groups. We also conducted literature reviews regarding regulations.

One of the key preliminary findings is the contrary regulative differences between South Korea and Japan. After the fabrication of Hwang Woo-suk's stem cell cloning and unethical human egg collection, bioethics law has been discussed for revision and the government seeks more strict regulation towards life science and healthcare. Then, the government takes strict positions towards bioindustries which provide genetic tests for children directly to the public or via clinics. The Korean National Bioethics Advisory Commission suggested the first genetic testing guidelines in 2006, which suggests banning certain types of genetic tests for multifactorial common diseases with poor scientific evidences. Finally, the Ministry banned 20 types of genetic tests to be supplied for the clinical and business purposes, which had been previously supplied by Korean bioindustries. The banned genetics tests include obesity, all cancers, Alzheimer's, high blood pressures, diabetes, osteoporosis, and asthma. Genetic testing for research purposes are out of scopes of these policies. South Korean bioethics law will be revised within a year or two and add statutes for regulating genetic testing. With regards to laboratory quality assurance, the Ministry of Health and Welfare requires all genetic testing labs in South Korea to submit applications about the list of genetic tests they provide and other information to qualify their laboratory quality to obtain licenses from the Ministry. The licensed labs have to accept laboratory inspection by the Korean Institute of Genetic Testing Evaluation (KIGTE).

On the contrary, in Japan, we don't have any legal regulation regarding genetic information and genetic tests. Ministry of Health and Welfare (MHLW) hasn't taken any action towards genetic testing markets and leave self-regulation by bioindustries. The Council for Protection of

Individual Genetic Information (CPIGI), which has been consisted of 25 bioindustries, published their first guidelines for standardization and business ethics on genetic tests marketing in 2007, supported by the Ministry of International Trade and Industry (MITI). The MHLW has made metabolic syndrome countermeasures and regular health check-ups for employees over 40 will be changed to stratify individual risks for metabolic syndromes since April 2008. It is obvious that bioindustries are ready to enter into the fast-growing market with genetic testing for obesity, which is banned in South Korea for clinical use. We could further compare the social and cultural difference between these neighbors, such as notions of individual responsibility and individuals' right to obtain genetic information for their healthcare management. We're now studying Taiwan and China's political approach for genetic testing business to obtain wider implications in East Asia. Some part of this project has been financially supported by the Japan Science and Technology Agency (JST) and the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

## 2. Ethical, legal and social support for "BioBank Japan" project

For supporting "BioBank Japan" project, led by Professor Yusuke Nakamura of Laboratory of Molecular Medicine of IMSUT, we've conducted three types of surveys and issued newsletters for participants. By the end of 2007, the project has obtained 200,000 written consent forms by research coordinators called Medical Coordinators (MC). The project trained nurses or pharmacists as MCs for obtaining fully informed consent from participants. This consent process had been well-worked out in advance and is complied with the government ethical guidelines for genetic/genomic research. However, recent publications show that the long and tedious consent process may not contribute to participants' understanding the overview of the research, may be unethical rather than ethical. If we long for "personalized medicine", we should think further about the construction of "personalized consent process" and we have to change the relationship between participants and researchers, from one-time informed consent to long lasting public trust.

Obtaining feedbacks from participants is also effective to keep incentives to participation and prevent dropout of participants from research process. We conducted three kinds of surveys to evaluate and improve the consent process and explore what the project should do for public involvement; questionnaire surveys towards research participants, a web-based questionnaire survey towards all MCs and focus group interviews with chief MCs to triangulate the consent process. The preliminary results show that participants are basically satisfied with the consent process and highly evaluate MCs' attitudes towards them. Most MCs also responded that they have made their original efforts to make their explanation easier and understandable specifically towards the elderly. However, certain amounts of participants have already forgotten about what for they have donated their DNA and serums and the experience of watching the DVD or the leaflet about the project overview. MCs explains that this project doesn't have any plans to disclose personal genotyped data to each participant, but a certain amount of participants responded that they now want to see their own genotyped data or tentative research feedbacks, while others are just satisfied with their contribution to genomic research without any rewards. Even though participants should forget the fact that they gave consent for research, MCs explain, encourage and appreciate participants at each time and participants recall their will for contribution.

To appreciate participants' and MCs' contribution to the project, we had issued "BioBank newsletters No.1" in August 2006 and "No.2" in November 2006. We will explore more methods and opportunities to communicate with participants. Because the current forms of BioBank newsletters are available only for the sighted with good eyesight, we make efforts for personalized information security to meet with disabilities of participants.

## Publications


1. 武藤香織．遺伝病ピアサポートの可能性と課題．遺伝診療をとりまく社会―その科学的・倫理的アプローチ．水口修紀・吉田雅幸監修，吉田雅幸・小笹由香編集．ブレーン出版．149–158，2007.
2. 武藤香織．神経難病当事者団体のネットワーキング．神経難病のすべて～症状・診断から最先端治療，福祉の実態まで～．阿部康二編著．新興医学出版社．187–192，2007.
3. 武藤香織．「オーダーメイド医療」からみえる科学性と不確実性．遺伝子技術の社会学．柘植あづみ・加藤秀一編著．文化書房博文社．177–181，2007.
4. 柊中智恵子，武藤香織．遺伝に対する相談へ

の対応．難病医療専門員による難病患者のための難病相談ガイドブック．吉良潤一編．九州大学出版会．64–87，2008.
5. Izumi Ishiyama, Akiko Nagai, Kaori Muto, Akiko Tamakoshi, Minori Kokado, Kyoko Mimura, Tetsuro Tanzawa, Zentaro Yamagata. Relationship between Public Attitudes toward Genomic Studies Related to Medicine and Their Level of Genomic Literacy in Japan. American Journal of Medical Genetics, 2008 (in press).
6. Kaori Muto; Truth telling and predictive genetic testing—Huntington's Disease in Japan, In Predictive & Genetic Testing in Asia: social-science perspectives on the ramification of choice. (University of Amsterdam Press, The Netherlands), 2008 (in press).