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(委嘱)	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

Since the completion of the Human Genome Project, high-throughput experimental projects have been initiated for uncovering genomic information in an extended sense, including transcriptomics, proteomics metabolomics, glycomics, and chemical genomics. We are developing a new generation of databases and computational technologies, beyond the traditional genome databases and sequence analysis tools, for making full use of these divergent and ever-increasing amounts of data, especially for medical and pharmaceutical applications.

1. KEGG DRUG and KEGG DISEASE

Minoru Kanehisa

KEGG is a database of biological systems that integrates genomic, chemical, and systemic functional information. It is widely used as a reference knowledge base for understanding higher-order functions and utilities of the cell 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/), a chemical structure based information resource for all approved drugs in the world, integrating target information in the context of KEGG pathways, efficacy information in the context of hierarchical drug classifications, and natural product information in the context of plant and other genomes. We also develop KEGG DISEASE (http://www.genome.jp/kegg/disease/), a new resource for understanding molecular mechanisms of human diseases, where molecular networks involving disease genes are represented as KEGG pathway maps or, when such details are not known, simply by a list of diseases genes and other lists of molecules including environmental factors.

2. KEGG OC: 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 4.2 million genes from over 1,000 organisms as of February 2009. 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. We built a system that performs automatic update of the ortholog cluster, which can be run weekly basis. As a result, we obtained 669,043 clusters including 403,752 singleton clusters from 3,721,464 protein coding genes. Among them, only 2,309 clusters were shared across kingdoms and other clusters were kingdom specific. The automatic classification of our ortholog clusters largely consistent with the manually curated ORTHOLOGY database. A web interface to search and browse genes in clusters is made available at http://oc.kegg.jp/.

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

Shuichi Kawashima, Yuki Moriya, Toshiaki Tokimatsu, Susumu Goto and Minoru Kanehisa

EGENES is a knowledge-based database for efficient analysis of plant expressed sequence tags (ESTs), which was recently added to the KEGG PLANT. It links plant genomic information with higher order functional information in KEGG. The genomic information in EGENES is a collection of EST contigs constructed from assembled plant ESTs by using EGassembler. The EST indices are automatically annotated with the KEGG Orthology identifiers (K numbers) by KEGG Automatics Annotation Server (KAAS). Currently, It contains 2,452,094 sequence catalogues (779,490 contigs and 1,672,604 singletons) in 62 plants. 25% of the sequences are assigned K numbers. EGENES is available at http://www.genome.jp/kegg/plant/pln_list.html.

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

Shuichi Kawashima, Toshiaki Katayama and Minoru Kanehisa

KEGG is a suite of databases and associated software, integrating our current knowledge of molecular interaction/reaction pathways and other systemic functions (PATHWAY and BRITE databases), 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 programmatically, we have been developing and maintaining the KEGG API as a stable SOAP/WSDL based web service. The KEGG API is available at http://www.genome.jp/kegg/soap/.

5. KEGG DAS: 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 bacterial organisms in the GENOME database in KEGG. Currently, KEGG DAS contains over 8 million annotations assigned to the genome sequences of 817 organisms (increased from 615 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. The KEGG DAS service is available at http://das.hgc.jp/.

6. Full-Arthropods: Constructing full length cDNA of pathogenetic arthropods

Toshiaki Katayama, Shuichi Kawashima, Miho Usui, Hiroyuki Wakaguri, Eri Kibukawa, Masahide Sasaki, Kazuhisa Hiranuka, Ryuichiro Maeda, Yutaka Suzuki, Sumio Sugano and Junichi Watanabe

Anopheles mosquito, tsetse fly, tsutsugamushi -mite, 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. Full-Arthropods is available on the site http://full-larth.hgc.jp/

7. Full-Entamoeba: a database for the full length cDNA library of Entamoeba

Toshiaki Katayama, Kazushi Hiranuka, Masahiro Kumagai, Yutaka Suzuki, Sumio Sugano, Atsushi Toyoda, Asao Makioka, Junichi Watanabe

Entamoeba histolytica is a protozoan parasite which predominantly infects humans and other primates and causes amebiasis. *E. histolytica* is estimated to infect about 50 million people worldwide and amebiasis is estimated to cause 70,000 deaths per year. Full-Entamoeba, a database for full-length cDNAs from a human parasite, *E. histolytica*, and a reptilian parasite, *E. invadens*, has been produced. The full-length cDNA libraries were produced using the oligo-capping method from trophozoites of each species cultivated axenically. A total of 5,000 5'end-one-pass sequences of cDNAs from the two species were compared with the non-redundant database of DDBL/GenBank/EMBL using BLAST and TBLASTX programs. These clones are available for further analysis and experiments. Full-Entamoeba database is available at http://full-lent.hgc.jp/

8. Analysis of sequence catalogs of the house dust mite *Dermatophagoides farinae*

Shuichi Kawashima, Atsushi Toyoda, Junichi Watanabe, Sadao Nogami and Minoru Kanehisa

The house dust mite is a cosmopolitan guest in human habitation and the multicellular organism that is one of the most closely associated with our life. It is now well established that the dust mites are major allergens causing bronchial asthma, allergic rhinitis and atopic dermatitis. *Dermatophagoides farinae* (American house dust mite) and *D. pteronissinus* (European house dust mite) are two most common species in the temperate zone. We produced the cDNA libraries of our *D. farinae* sample containing young nymphs and adults using the vector trapper method and sequenced the both ends of 11,520 cDNA clones. Cleaning, clustering and assembling of the row sequences produced 3,031 contigs and 4,281 singletons. 1,797 of the total unique 7,312 sequences were assigned KEGG Orthology by KAAS system. More than 30% of the sequences showed significant matches to KEGG GENES database, which includes well characterized Der f group 1 allergens. We predicted 1,109 peptides longer than 20 amino acids from the 3,031 contigs. Some of the peptides are predicted to contain the 9-mer peptides with strong affinities to MHC class II by NetMHC 3.0. We expect that these *in silico* analyses will pave the way toward prediction of allergens from *D. farinae*.

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

Toshiaki Katayama, Shuichi Kawashima, Kazuhiro Ohi, 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 98,868,465 (Release 169+ daily updates). In order for such a vast amount of data to be searched at a high speed, we have developed a high performance database entry retrieval system, named HiGet. For this purpose, the system is constructed on the HiRDB, a commercial ORDBMS (Object-oriented Relational Database Management System) developed by Hitachi, Ltd. HiGet can perform full text search on various biological databases including GenBank, RefSeq, UniProt, Prosite, OMIM and PDB. Additional advantage of the HiGet system is the capability of a field specific search, which enables users to narrow down the number of results, especially useful for collecting sequences of their specific needs. We have also developed a sequence similarity search (SSS) service to find homologous sequences with various algorithms including BLAST, FASTA, SSEARCH, TRANS, and EXONERATE. This variety of options is unique among the public services and users can select an appropriate method to search similar sequences according to their query. Because algorithms such as TRANS and EXONERATE are highly time consuming, the SSS service is backended by the distributed computing environment with the Sun Grid Engine in our super computer system. HiGet and SSS services are available at http://higet.hgc.jp/ and http://sss.hgc.jp/ respectively.

10. Linear-time protein 3-D structure searching algorithm

Tetsuo Shibuya

Finding similar structures from 3-D structure databases of proteins (or other molecules) is becoming one of the most important issues in the post-genomic molecular biology. To compare 3-D structures of two molecules, biologists mostly use the RMSD (root mean square deviation) as the similarity measure. The RMSD is one of the most fundamental similarity measures used in various fields, such as computer vision and robotics, for comparing two sets of coordinates. In this research, we propose new theoretically and practically fast algorithms for the basic problem of finding all the substructures of structures in a structure database of chain molecules (such as proteins), whose RMSDs to the query are within a given constant threshold. The best-known worst-case time complexity for the problem is $O(N \log m)$, where N is the database size and m is the query size. The previous best-known expected time complexity for the problem is also $O(N \log m)$. In this research, we propose a new breakthrough linear-expected-time algorithm. It is not only a theoretically significant improvement over previous algorithms, but also a practically faster algorithm, according to computational experiments. We also propose a series of preprocessing algorithms that enable faster queries, though there have been no known indexing algorithm whose query time complexity is better than the above $O(N \log m)$ bound. One is an $O(N \log^2 N)$-time and $O(N \log N)$-space preprocessing algorithm with expected query time complexity of $O(m+N/m^{0.5})$. Another is an $O(N \log N)$-time and $O(N)$-space preprocessing algorithm with expected query time complexity of $O(N/m^{0.5}+m \log (N/m))$.

We also extend the above linear-time algorithm into an algorithm with expected query time complexity of $O(m+N/m^{1-\varepsilon})$, where ε is an arbitrary small constant such that $0<\varepsilon<1$. We furthermore extend the above linear-time algorithm so that it can deal with insertions and deletions.

We checked the performance of our linear-expected-time algorithm through computational experiments over the whole PDB database. The experiments show that our algorithm is much faster than the previous algorithms. For example, our algorithm is 3.6 to 28 times faster than previously known algorithms, to search for similar substructures whose RMSDs are within 1 Å to queries of ordinary lengths. The experiments also show that there is consistency between the above theoretical results and the experimental results. In other words, the actual computation time of our linear-expected-time algorithm is not influenced by the difference of query lengths, in contrast to previous algorithms.

11. 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 research, 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.

12. Fast flexible protein structure alignment

Kohichi Suematsu and Tetsuo Shibuya

The Hinge Detection Algorithm described in section 11 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 ex-

ponential time. For the algorithm, we developed a detailed score matrix for comparing local structures based on the naïve Bayse learning.

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. 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. We developed a lazy algorithm based on the Schurmann-Stoye algorithm, which is more efficient than both Boyer-Moore algorithm and other suffix tree-based algorithms in case the number of queries is limited.

15. Color space-DNA sequence mapping alignment algorithm

Ben Hachimori and Tetsuo Shibuya

Applied Biosystems's SOLiD system encode the DNA sequence into a sequence of data type called the color space, where one of 4 fluorescent colors is assigned to each two adjacent base's 16 pattern orderings. However, there have been known no algorithm that aligns/maps the color-space sequence to the DNA sequence with consideration of the difference between the experimental error and the actual mutation. We developed an alignment algorithm that distinguishes the experimental error and actual DNA mutation to align the color-space data against ordinary DNA sequences. Moreover, we computed the optimal score table for the alignment based on the actual *E. coli* data.

16. 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.

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: D480-D484, 2008.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research* 36: D202-D 205, 2008.

Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S. and Kanehisa, M. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Research* 36: W423-426, 2008.

Wakaguri, H., Suzuki, Y., Katayama, T., Kawashima, S., Kibukawa, E., Hiranuka, K., Sasaki, M., Sugano, S. and Watanabe, J. Full-Malaria/Parasites and Full-Arthropods: database of full-length cDNAs of parasites and arthropods, update 2009. *Nucleic Acids Research*

37: D520-D525, 2008.
Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W. and Kanehisa, M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24: i232-i240, 2008.
Takarabe, M., Okuda, S., Itoh, M., Tokimatsu, T., Goto, S. and Kanehisa, M. Network analysis of adverse drug interactions. *Genome Informatics* 20: 252-259, 2008.
Hashimoto, K., Yoshizawa, A.C., Okuda, S., Kuma, K., Goto, S. and Kanehisa, M. The repertoire of desaturases and elongases reveals fatty acid variations in 56 eukaryotic genomes. *J. Lipid Res*. 49: 183-191 (2008).
Shibuya, T., Fast Hinge Detection Algorithms for Flexible Protein Structures, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, to appear.
Shibuya, T., Searching Protein 3-D Structures in Linear Time, *Proc. 13th Annual International Conference on Research in Computational Molecular Biology* (*RECOMB 2009*), 2009, to appear.
Shibuya, T., Linear-Time Algorithm for Searching Protein 3-D Structures, *IPSJ SIG Notes SIGAL 123-4*, 2009, to appear.
Suematsu, K., Shibuya, T., Flexible Protein Alignment of 3D-Structures Allowing Dynamic Transformation, *ISPSJ SIG Notes SIGBIO 12-12*, 2008, pp. 87-94.
本多渉，田辺麻央，矢野亜津子，金久實．バイオインフォマティクス・システムバイオロジーとKEGG，生化学，80⑿：1094−1111，2008

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
Project AssistantProfessor	Yoshinori Tamada, Ph.D.	特任助教	博士(情報学)	玉　田　嘉　紀

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. Systematic reconstruction of TRANSPATH data into Cell System Markup Language

Masao Nagasaki, Ayumu Saito, Chen Li, Euna Jeong, Satoru Miyano

Many biological repositories store information based on experimental study of the biological processes within a cell, such as protein-protein interactions, metabolic pathways, signal transduction pathways, or regulations of transcription factors and miRNA. Unfortunately, it is difficult to directly use such information when generating simulation-based models. Thus, modeling rules for encoding biological knowledge into system-dynamics-oriented standardized formats would be very useful for full understanding of cellular dynamics at the system level. We selected the TRANSPATH database, a manually curated high-quality pathway database, which provides a rich source of cellular events in humans, mice, and rats, curated from over 31,500 papers. In this work, we defined 16 modeling rules based on hybrid functional Petri net with extension (HFPNe), which is suitable for graphical representation and simulation of biological processes. In these modeling rules, each Petri net element is incorporated with Cell System Ontology (CSO) to enable semantic interoperability of models. As a formal ontology for biological pathway modeling with dynamics, CSO also defines biological terminology and corresponding icons. By combining HFPNe with the CSO features, we made a method for transform ing TRANSPATH data to simulation-based semantically valid models. The results are encoded into a biological pathway format, Cell System Markup Language (CSML), which eases the exchange and integration of biological data and models. By using the 16 modeling rules, 97% of the reactions in TRANSPATH are converted into simulation-based models represented in CSML. This reconstruction demonstrated that it is possible to use our rules to generate quantitative models from static pathway descriptions.

b. Finding optimal Bayesian network given a super-structure

Eric Perrier, Seiya Imoto, Satoru Miyano

Conventional approaches for learning Bayesian network structure from data have disadvantages in terms of complexity and lower accuracy of their results. However, a recent empirical study has shown that a hybrid algorithm improves sensitively accuracy and speed: it learns a skeleton with an independency test (IT) approach and constrains on the directed acyclic graphs considered during the search-and-score phase. Subsequently, we defined the structural constraint by introducing the concept of super-structure S, which is an undirected graph that restricts the search to networks whose skeleton is a subgraph of S. We developed a super-structure constrained optimal search (COS): its time complexity is upper bounded by $O(\gamma_m^n)$, where $\gamma_m < 2$ depends on the maximal degree m of S. Empirically, complexity depends on the average degree m' and sparse structures allow larger graphs to be calculated. Our algorithm is faster than an optimal search by several orders and even finds more accurate results when given a sound super-structure. Practically, S can be approximated by IT approaches; significance level of the tests controls its sparseness, enabling to control the trade-off between speed and accuracy. For incomplete super-structures, a greedily post-processed version (COS+) still enables to significantly outperform other heuristic searches.

c. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models

Osamu Hirose, Ryo Yoshida[1], Seiya Imoto, Rui Yamaguchi, Tomoyuki Higuchi[1], D. Stephen Charnock-Jones[2], Cristin Print[3], Satoru Miyano; [1]Institute of Statistical Mathematics, [2]Cambridge University, [3]University of Auckland

We developed a novel method based on the state space model to identify the transcriptional modules and module-based gene networks simultaneously. The state space model has the potential to infer large-scale gene networks, e.g. of order 10^3, from time-course gene expression profiles. Particularly, we succeeded in identification of a cell cycle system by using the gene expression profiles of *Saccharomyces cerevisiae* in which the length of the time-course and number of genes were 24 and 4382, respectively. However, when analyzing shorter time-course data, e.g. of length 10 or less, the parameter estimations of the state space model often fail due to overfitting. To extend the applicability of the state space model, we provided an approach to use the technical replicates of gene expression profiles, which are often measured in duplicate or triplicate. The use of technical replicates is important for achieving highly-efficient inference of gene networks with short time-course data. The potential of the proposed method were demonstrated through the time-course analysis of the gene expression profiles of human umbilical vein endothelial cells undergoing growth factor deprivation-induced apoptosis.

d. Predicting differences in gene regulatory systems by state space models

Rui Yamaguchi, Seiya Imoto, Mai Yamauchi, Masao Nagasaki, Ryo Yoshida[1], Teppei Shimamura, Yosuke Hatanaka, Kazuko Ueno, Tomoyuki Higuchi[1], Noriko Gotoh, Satoru Miyano

We developed a statistical method to predict differentially regulated genes of case and control samples from time-course gene expression data by leveraging unpredictability of the expression patterns from the underlying regulatory system inferred by a state space model. The proposed method can screen out genes that show different patterns but generated by the same regulations in both samples, since these patterns can be predicted by the same model. Our strategy consists of three steps. Firstly, a gene regulatory system is inferred from the control data by a state space model. Then the obtained model for the underlying regulatory system of the control sample is used to predict the case data. Finally, by assessing the significance of the difference between case and predicted-case time-course data of each gene, we are able to detect the unpredictable genes that are the candidate as the key differences between the regulatory systems of case and control cells. We illustrate the whole process of the strategy by an actual example, where human small airway epithelial cell gene regulatory systems were generated from novel time courses of gene expressions following treatment with(case)/without(control) the drug gefitinib, an inhibitor for the epidermal growth factor receptor tyrosine kinase. Finally, in gefitinib response data we succeeded in finding unpredictable genes that are candidates of the specific targets of gefitinib. We also discussed differences in regulatory systems for the unpredictable genes. The proposed method would be a promising tool for identifying biomarkers and drug target genes.

e. Bayesian learning of biological pathways on genomic data assimilation

Ryo Yoshida[1], Masao Nagasaki, Rui Yamaguchi, Seiya Imoto, Satoru Miyano, Tomoyuki Higuchi[1]

Mathematical modeling and simulation, based on biochemical rate equations, provide us a rigorous tool for unraveling complex mechanisms of biological pathways. To proceed to simulation experiments, it is an essential first step to find effective values of model parameters, which are difficult to measure from *in vivo* and *in vitro* experiments. Furthermore, once a set of hypothetical models has been created, any statistical criterion is needed to test the ability of the constructed models and to proceed to model revision. We developed a new statistical technology towards data-driven construction of *in silico* biological pathways. The method starts with a knowledge-based modeling with hybrid functional Petri net. It then proceeds to the Bayesian learning of model parameters for which experimental data are available. This process exploits quantitative measurements of evolving biochemical reactions, e.g. gene expression data. Another important issue that we consider is statistical evaluation and comparison of the constructed hypothetical pathways. For this purpose, we have developed a new Bayesian information-theoretic measure that assesses the predictability and the biological robustness of *in silico* pathways.

f. Modeling nonlinear gene regulatory networks from time series gene expression data

André Fujita, João Ricardo Sato[5], Humberto Miguel Garay-Malpartida[5], Mari Cleide Sogayar[5], Carlow Eduardo Ferreira[5], Satoru Miyano; [5]University of São Paulo

In cells, molecular networks such as gene regulatory networks are the basis of biological complexity. Therefore, gene regulatory networks have become the core of research in systems biology. Understanding the processes underlying the several extracellular regulators, signal transduction, protein-protein interactions, and differential gene expression processes requires detailed molecular description of the protein and gene networks involved. To understand better these complex molecular networks and to infer new regulatory associations, we developed a statistical method based on vector autoregressive models and Granger causality to estimate nonlinear gene regulatory networks from time series microarray data. Most of the models available in the literature assume linearity in the inference of gene connections; moreover, these models do not infer directionality in these connections. Thus, a priori biological knowledge is required. However, in pathological cases, no a priori biological information is available. To overcome these problems, we present the nonlinear vector autoregressive (NVAR) model. We have applied the NVAR model to estimate nonlinear gene regulatory networks based entirely on gene expression profiles obtained from DNA microarray experiments. We showed the results obtained by NVAR through several simulations and by the construction of three actual gene regulatory networks (p53, NF-κB, and c-Myc) for HeLa cells.

g. Fast grid layout algorithm for biological networks with sweep calculation

Kaname Kojima, Masao Nagasaki, Satoru Miyano

Properly drawn biological networks are of great help in the comprehension of their characteristics. The quality of the layouts for retrieved biological networks is critical for pathway databases. However, since it is unrealistic to manually draw biological networks for every retrieval, automatic drawing algorithms are essential. Grid layout algorithms handle various biological properties such as aligning vertices having the same attributes and complicated positional constraints according to their subcellular localizations; thus, they succeed in providing biologically comprehensible layouts. However, existing grid layout algorithms are not suitable for real-time drawing, which is one of requisites for applications to pathway databases, due to their high-computational cost. In addition, they do not consider edge directions and their resulting layouts lack traceability for biochemical reactions and gene regulations, which are the most important features in biological networks. We devised a new calculation method termed sweep calculation and reduced the time complexity of the current grid layout algorithms through its encoding and decoding processes. We conduct ed practical experiments by using 95 pathway models of various sizes from TRANSPATH and showed that our new grid layout algorithm is much faster than existing grid layout algorithms. For the cost function, we introduced a new component that penalizes undesirable edge directions to avoid the lack of traceability in pathways due to the differences in direction between in-edges and out-edges of each vertex.

h. Estimation of nonlinear gene regulatory networks via L_1 regularized NVAR from time series gene expression data

Kaname Kojima, André Fujita, Teppei Shimamura, Seiya Imoto, Satoru Miyano

Recently, nonlinear vector autoregressive (NVAR) model based on Granger causality was proposed to infer nonlinear gene regulatory networks from time series gene expression data. Since NVAR requires a large number of parameters due to the basis expansion, the length of time series microarray data is insufficient for accurate parameter estimation and we need to limit the size of the gene set strongly. To address this limitation, we employed L_1 regularization technique to estimate NVAR. Under L_1 regularization, direct parents of each gene can be selected efficiently even when the number of parameters exceeds the number of data samples. We can thus estimate larger gene regulatory networks more accurately than those from existing methods. Through the simulation study, we verified the effectiveness of the proposed method by comparing its limitation in the number of genes to that of the existing NVAR. The proposed method was also applied to time series microarray data of Human hela cell cycle.

i. Multivariate gene expression analysis reveals functional connectivity changes between normal/tumoral prostates

André Fujita, Luciana Rodrigues Gomes[5], João Ricardo Sato[6], Rui Yamaguchi, Carlos Eduardo Thomaz[7], Mari Cleide Sogayar[5], Satoru Miyano; [6]Universidade Federal do ABC, [7]Centro Universitário da FEI

Principal Component Analysis (PCA) combined with the Maximum-entropy Linear Discriminant Analysis (MLDA) was applied in order to identify genes with the most discriminative information between normal and tumoral prostatic tissues. Data analysis was carried out using three different approaches, namely: (i) differences in gene expression levels between normal and tumoral conditions from a univariate point of view; (ii) in a multivariate fashion using MLDA; and (iii) with a dependence network approach. Our results show that malignant transformation in the prostatic tissue is more related to functional connectivity changes in their dependence networks than to differential gene expression. The MYLK, KLK2, KLK3, HAN11, LTF, CSRP1 and TGM4 genes presented significant changes in their functional connectivity between normal and tumoral conditions and were also classified as the top seven most informative genes for the prostate cancer genesis process by our discriminant analysis. Moreover, among the identified genes we found classically known biomarkers and genes which are closely related to tumoral prostate, such as KLK3 and KLK2 and several other potential ones. We have demonstrated that changes in functional connectivity may be implicit in the biological process which renders some genes more informative to discriminate between normal and tumoral conditions. Using the proposed method, namely, MLDA, in order to analyze the multivariate characteristic of genes, it was possible to capture the changes in dependence networks which are related to cell transformation.

j. Rule-based reasoning for system dynamics in cell systems

Euna Jeong, Masao Nagasaki, Satoru Miyano

A system-dynamics-centered ontology, called the Cell System Ontology (CSO), has been developed for representation of diverse biological pathways. Many of the pathway data based on the ontology have been created from databases via data conversion or curated by expert biologists. It is essential to validate the pathway data which may cause unexpected issues such as semantic inconsistency and incompleteness. This paper discusses three criteria for validating the pathway data based on CSO as follows: (1) structurally correct models in terms of Petri nets, (2) biologically correct models to capture biological meaning, and (3) systematically correct models to reflect biological behaviors. Simultaneously, we have investigated how logic-based rules can be used for the ontology to extend its expressiveness and to complement the ontology by reasoning, which aims at qualifying pathway knowledge. Finally, we show how the proposed approach helps exploring dynamic modeling and simulation tasks without prior knowledge.

k. A novel strategy to search conserved transcription factor binding sites among coexpressing genes in human

Yosuke Hatanaka, Masao Nagasaki, Rui Yamaguchi, Takeshi Obayashi, Kazuyuki Numata, André Fujita, Teppei Shimamura, Yoshinori Tamada, Seiya Imoto, Kengo Kinoshita, Kenta Nakai, Satoru Miyano

We reported various transcription factor binding sites (TFBSs) conserved among co-expressed genes in human promoter region using expres-

sion and genomic data. Assuming similar promoter structure induces similar transcriptional regulation, hence induces similar expression profile, we compared the promoter structure similarities between co-expressed genes. Comprehensive TF binding site predictions for all human genes were conducted for 19,777 promoter regions around the transcription start site (TSS) given from DBTSS and promoter similarity search were conducted among coexpressing genes data provided from newly developed COXPRESdb. Combination of Position Weight Matrix (PWM) motif prediction and bootstrap method, 7,313 genes have at least one statistically significant conserved TFBS. We also applied basket method analysis for seeking combinatorial activities of those conserved TFBSs.

l. Simulation analysis for the effect of light-dark cycle on the entrainment in circadian rhythm

Natumi Mitou[8], Yuto Ikegami[8], Hiroshi Matsuno[8], Satoru Miyano, Shin-ichi T. Inouye[8]; [8]Yamaguchi University

Circadian rhythms of the living organisms are 24hr oscillations found in behavior, biochemistry and physiology. Under constant conditions, the rhythms continue with their intrinsic period length, which are rarely exact 24hr. In this paper, we examine the effects of light on the phase of the gene expression rhythms derived from the interacting feedback network of a few clock genes, taking advantage of a computer simulation with Cell Illustrator. The simulation results suggested that the interacting circadian feedback network at the molecular level is essential for phase dependence of the light effects, observed in mammalian behavior. Furthermore, the simulation reproduced the biological observations that the range of entrainment to shorter or longer than 24hr light-dark cycles is limited, centering around 24hr. Application of our model to inter-time zone flight successfully demonstrated that 6 to 7 days are required to recover from jet lag when traveling from Tokyo to New York.

2. Statistical and Computational Knowledge Discovery

a. Nonlinear regression modeling via regularized radial basis function networks

Tomohiro Ando[9], Sadanori Konishi[10], Seiya Imoto; [9]Keio University, [10]Kyushu University

The problem of constructing nonlinear regression models is investigated to analyze data with complex structure. We introduced radial basis functions with hyperparameter that adjusts the amount of overlapping basis functions and adopts the information of the input and response variables. By using the radial basis functions, we constructed nonlinear regression models with help of the technique of regularization. Crucial issues in the model building process are the choices of a hyperparameter, the number of basis functions and a smoothing parameter. We present information-theoretic criteria for evaluating statistical models under model misspecification both for distributional and structural assumptions. We used real data examples and Monte Carlo simulations to investigate the properties of the proposed nonlinear regression modeling techniques. The simulation results showed that our nonlinear modeling performs well in various situations, and clear improvements were obtained for the use of the hyperparameter in the basis functions.

b. The GC and window-averaged DNA curvature profile of secondary metabolite gene cluster in *Aspergillus fumigatus* genome

Jin Hwan Do, Satoru Miyano

An immense variety of complex secondary metabolites is produced by filamentous fungi including *Aspergillus fumigatus*, a main inducer of invasive aspergillosis. The identification of fungal secondary metabolite gene cluster is essential for the characterization of fungal secondary metabolism in terms of genetics and biochemistry through recombinant technologies such as gene disruption and cloning. Most of the prediction methods for secondary metabolite gene cluster severely depend on homology searches. However, homology-based approach has intrinsic limitation to unknown or novel gene cluster. We analyzed the GC and window-averaged DNA curvature profile of 26 secondary metabolite gene clusters in the *A. fumigatus* genome to find out potential conserved features of secondary metabolite gene cluster. Fifteen secondary metabolite gene clusters showed a conserved pattern in window-averaged DNA curvature profile, that is, the DNA regions including secondary metabolic signature genes such as polyketide synthase, nonribosomal peptide synthase, and/or dimethylallyl tryptophan synthase consisted of window-averaged DNA curvature values lower than 0.18 and these DNA regions were at least 20 kb. Forty percent of secondary metabolite gene clusters with this conserved pattern were related to severe regulation by a transcription factor, LaeA. Our result could be used

for identification of other fungal secondary metabolite gene clusters, especially for secondary metabolite gene cluster that is severely regulated by LaeA or other proteins with similar function to LaeA.

c. ExonMiner: Web service for analysis of GeneChip exon array data

Kazuyuki Numata, Ryo Yoshida[1], Masao Nagasaki, Ayumu Saito, Seiya Imoto, Satoru Miyano

Some splicing isoform-specific transcriptional regulations are related to disease. Therefore, detection of disease specific splice variations is the first step for finding disease specific transcriptional regulations. Affymetrix Human Exon 1.0 ST Array can measure exon-level expression profiles that are suitable to find differentially expressed exons in genome-wide scale. However, exon array produces massive datasets that are more than we can handle and analyze on personal computer. We have developed ExonMiner that is the first all-in-one web service for analysis of exon array data to detect transcripts that have significantly different splicing patterns in two cells, e.g. normal and cancer cells. ExonMiner can perform the following analyses: (1) data normalization, (2) statistical analysis based on two-way ANOVA, (3) finding transcripts with significantly different splice patterns, (4) efficient visualization based on heatmaps and barplots, and (5) meta-analysis to detect exon level biomarkers. We implemented ExonMiner on the supercomputer system of Human Genome Center in order to perform genome-wide analysis for more than 300,000 transcripts in exon array data, which has the potential to reveal the aberrant splice variations in cancer cells as exon level biomarkers. ExonMiner is well suited for analysis of exon array data and does not require any installation of software except for internet browsers. The URL of ExonMiner is http://ae.hgc.jp/exonminer. Users can analyze full dataset of exon array data within hours by high-level statistical analysis with sound theoretical basis that finds aberrant splice variants as biomarkers.

Publications

1. Ando, T., Konishi, S., Imoto, S. Nonlinear regression modeling via regularized radial basis function networks. Journal of Statistical Planning and Inference. 138 (11): 3616-3633, 2008.
2. Brazma, A., Miyano, S., Akutsu, T. Proceedings of the 6th Asia-Pacific Bioinformatics Conference (APBC 2008). Imperial College Press, 2008.
3. Do, J.H., Miyano, S. The GC and window-averaged DNA curvature profile of secondary metabolite gene cluster in *Aspergillus fumigatus* genome. Applied Microbiology and Biotechnology. 80 (5): 841-847, 2008.
4. Fujita, A., Gomes, L.R., Sato, J.R., Yamaguchi, R., Thomaz, C.E., Sogayar, M.C., Miyano, S. Multivariate gene expression analysis reveals functional connectivity changes between normal/tumoral prostates. BMC Systems Biology. 2: 106, 2008.
5. 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. 6 (5): 961-979, 2008.
6. Hatanaka, Y., Nagasaki, M., Yamaguchi, R., Obayashi, T., Numata, K., Fujita, A., Shimamura, T., Tamada, Y., Imoto, S., Kinoshita, K., Nakai, K., Miyano, S. A novel strategy to search conserved transcription factor binding sites among coexpressing genes in human. Genome Informatics. 20: 212-221, 2008.
7. 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. 24 (7): 932-942, 2008.
8. Hirose, O., Yoshida, R., Yamaguchi, R., Imoto, S., Higuchi, T., Miyano, S. Analyzing time course gene expression data with biological and technical replicates to estimate gene networks by state space models. Proc. 2nd Asia International Conference on Modelling & Simulation, 940-946, 2008. (AMS 2008: Refereed conference)
9. Jeong, E., Nagasaki, M., Miyano, S. Rule-based reasoning for system dynamics in cell systems. Genome Informatics. 20: 25-36, 2008.
10. Kitakaze, H., Kanda, M., Nakatsuka, H., Ikeda, N., Matsuno, H., Miyano, S. Prediction of fragile points for robustness checking of cell systems. IEICE TRANSACTIONS on Information and Systems D. J91-D (9): 2404-2417, 2008.
11. Knapp, E.-W., Benson, G., Holzhutter, H.-G., Kanehisa, M., Miyano, S. (Eds.) Genome Informatics. 20, 2008.
12. Kojima, K., Fujita, A., Shimamura, T., Imoto,

S., Miyano, S. Estimation of nonlinear gene regulatory networks via L1 regularized NVAR from time series gene expression data. Genome Informatics. 20: 37-51, 2008.
13. Kojima, K., Nagasaki, M., Miyano, S. Fast grid layout algorithm for biological networks with sweep calculation. Bioinformatics. 24 (12): 1426-1432, 2008.
14. Mito, N., Ikegami, Y., Matsuno, H., Miyano, S., Inouye, S. Simulation analysis for the effect of light-dark cycle on the entrainment in circadian rhythm. Genome Informatics. 21: 212-223, 2008.
15. Nagasaki, M., Saito, A., Chen, L., Jeong, E., Miyano, S. Systematic reconstruction of TRANSPATH data into Cell System Markup Language. BMC Systems Biology. 2: 53, 2008.
16. Niida, A., Smith, A.D., Imoto, S., Tsutsumi, S., Aburatani, H., Zhang, M.Q., Akiyama, T. Integrative bioinformatics analysis of transcriptional regulatory programs in breast cancer cells. BMC Bioinformatics. 9: 404, 2008.
17. Numata, K., Yoshida, R., Nagasaki, M., Saito, S., Imoto, S., Miyano, S. ExonMiner: Web service for analysis of GeneChip exon array data. BMC Bioinformatics. 9: 494, 2008.
18. Numata, K., Imoto, S., Miyano, S. Partial order-based Bayesian network learning algorithm for estimating gene networks. Proc. IEEE 8th International Symposium on Bioinformatics & Bioengineering, IEEE Computer Society, 357-360, 2008. (BIBM 2008: Refereed conference)
19. Perrier, E., Imoto, S., Miyano, S. Finding optimal Bayesian network given a superstructure. J. Machine Learning Research. 9: 2251-2286, 2008.
20. Yamaguchi, R., Imoto, S., Yamauchi, M., Nagasaki, M., Yoshida, R., Shimamura, T., Hatanaka, Y., Ueno, K., Higuchi, T., Gotoh, N., Miyano, S. Predicting differences in gene regulatory systems by state space models. Genome Informatics. 21: 101-113, 2008.
21. Yoshida, R., Nagasaki, M., Yamaguchi, R., Imoto, S., Miyano, S., Higuchi, T. Bayesian learning of biological pathways on genomic data assimilation. Bioinformatics. 24(22): 2592-2601, 2008.

Human Genome Center

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

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

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

1. Genes playing significant roles in human cancer

Toyomasa Katagiri, Yataro Daigo, Hidewaki Nakagawa, Hitoshi Zembutsu, Koichi Matsuda, Ryuji Hamamoto, Sachiko Dobashi, Tomomi Ueki, Chikako Fukukawa, Eiji Hirota, Meng-Lay Lin, Jae-Hyun Park, Yosuke Harada, Satoshi Nagayama, Toshihiko Nishidate, Arata Shimo, Masahiko Ajiro, Jung-Won Kim, Tatsuya Kato, Daizaburo Hirata, Koji Ueda, Atsushi Takano, Nobuhisa Ishikawa, Koji Takahashi, Takumi Yamabuki, Nagato Sato, Nguyen Minh-Hue, Ryohei Nishino, Junkichi Koinuma, Daiki Miki, Ken Masuda, Masato Aragaki, Dragomira Nikolaeva Nikolova, Satoko Uno, Yoichiro Kato, Kenji Tamura, Kotoe Kashiwaya, Masayo Hosokawa, Shingo Ashida, Su-Youn Chung, Motohide Uemura, Lianhua Piao, Chizu Tanikawa, Motoko Unoki, Masanori Yoshimatsu, Shinya Hayami, and Yusuke Nakamura

(1) Lung cancer

DLX5 (distal-less homeobox 5)

We found that distal-less homeobox 5 (DLX5) gene, a member of the human distal-less homeobox transcriptional factor family was overexpressed in the great majority of lung cancers. Northern blot and immunohistochemical analyses detected expression of *DLX5* only in placenta among 23 normal tissues examined. Immunohistochemical analysis showed that positive immunostaining of DLX5 was correlated with tumor size (pT classification; $P = 0.0053$) and poorer prognosis of non-small cell lung can-

cer patients ($P=0.0045$). It was also shown to be an independent prognostic factor ($P=0.0415$). Treatment of lung cancer cells with small interfering RNAs for DLX5 effectively knocked down its expression and suppressed cell growth. These data implied that DLX5 is useful as a target for the development of anticancer drugs and cancer vaccines as well as for a prognostic biomarker in clinic.

ECT2 (epithelial cell transforming sequence 2)

We screened for genes that were frequently overexpressed in the tumors through gene expression profile analyses of 101 lung cancers and 19 esophageal squamous cell carcinomas (ESCC) by cDNA microarray consisting of 27,648 genes or expressed sequence tags. In this process, we identified epithelial cell transforming sequence 2 (ECT2) as a candidate. Northern blot and immunohistochemical analyses detected expression of ECT2 only in testis among 23 normal tissues. Immunohistochemical staining showed that a high level of ECT2 expression was associated with poor prognosis for patients with NSCLC ($P=0.0004$) as well as ESCC ($P=0.0088$). Multivariate analysis indicated it to be an independent prognostic factor for NSCLC ($P=0.0005$). Knockdown of ECT2 expression by small interfering RNAs effectively suppressed lung and esophageal cancer cell growth. In addition, induction of exogenous expression of ECT2 in mammalian cells promoted cellular invasive activity. ECT2 cancer-testis antigen is likely to be a prognostic biomarker in clinic and a potential therapeutic target for the development of anticancer drugs and cancer vaccines for lung and esophageal cancers.

(2) Breast Cancer

DTL/RAMP (denticleless/RA-regulated nuclear matrix associated protein)

To investigate the detailed molecular mechanism of mammary carcinogenesis and discover novel therapeutic targets, we previously analysed gene expression profiles of breast cancers. We here report characterization of a significant role of DTL/RAMP (denticleless/RA-regulated nuclear matrix associated protein) in mammary carcinogenesis. Semiquantitative RT-PCR and northern blot analyses confirmed upregulation of *DTL/RAMP* in the majority of breast cancer cases and all of breast cancer cell lines examined. Immunocytochemical and western blot analyses using anti-DTL/RAMP polyclonal antibody revealed cell-cycle-dependent localization of endogenous DTL/RAMP protein in breast cancer cells; nuclear localization was observed in cells at interphase and the protein was concentrated at the contractile ring in cytokinesis process. The expression level of DTL/RAMP protein became highest at G(1)/S phases, whereas its phosphorylation level was enhanced during mitotic phase. Treatment of breast cancer cells, T47 D and HBC4, with small-interfering RNAs against DTL/RAMP effectively suppressed its expression and caused accumulation of G(2)/M cells, resulting in growth inhibition of cancer cells. We further demonstrate the *in vitro* phosphorylation of DTL/RAMP through an interaction with the mitotic kinase, Aurora kinase-B (AURKB). Interestingly, depletion of AURKB expression with siRNA in breast cancer cells reduced the phosphorylation of DTL/RAMP and decreased the stability of DTL/RAMP protein. These findings imply important roles of DTL/RAMP in growth of breast cancer cells and suggest that DTL/RAMP might be a promising molecular target for treatment of breast cancer.

(3) Renal cancer

TMEM22 (transmembrane protein 22)

In order to clarify the molecular mechanism involved in renal carcinogenesis, and to identify molecular targets for development of novel treatments of renal cell carcinoma (RCC), we previously analyzed genome-wide gene expression profiles of clear-cell types of RCC by cDNA microarray. Among the transcativated genes, we herein focused on functional significance of TMEM22 (transmembrane protein 22), a transmembrane protein, in cell growth of RCC. Northern blot and semi-quantitative RT-PCR analyses confirmed up-regulation of *TMEM22* in a great majority of RCC clinical samples and cell lines examined. Immunocytochemical analysis validated its localization at the plasma membrane. We found an interaction between TMEM 22 and RAB37 (Ras-related protein Rab-37), which was also up-regulated in RCC cells. Interestingly, knockdown of either of TMEM22 or RAB37 expression by specific siRNA caused significant reduction of cancer cell growth. Our results imply that the TMEM22/RAB37 complex is likely to play a crucial role in growth of RCC and that inhibition of the TMEM22/RAB37 expression or their interaction should be novel therapeutic targets for RCC.

(4) Synovial sarcoma

FZD10 (Frizzled homologue 10)

We previously reported that Frizzled homologue 10 (FZD10), a member of the Wnt signal receptor family, was highly and specifically upregulated in synovial sarcoma and played critical roles in its cell survival and growth. We investigated a possible molecular mechanism of the FZD10 signaling in synovial sarcoma cells. We found a significant enhancement of phosphorylation of the Dishevelled (Dvl)2/Dvl3 complex as well as activation of the Rac1-JNK cascade in synovial sarcoma cells in which FZD 10 was overexpressed. Activation of the FZD10-Dvls-Rac1 pathway induced lamellipodia formation and enhanced anchorage-independent cell growth. FZD10 overexpression also caused the destruction of the actin cytoskeleton structure, probably through the downregulation of the RhoA activity. Our results have strongly implied that FZD10 transactivation causes the activation of the non-canonical Dvl-Rac1-JNK pathway and plays critical roles in the development/progression of synovial sarcomas.

(5) Pancreatic cancer

CST6 (Cystatin 6)

Pancreatic ductal adenocarcinoma (PDAC) shows the worst mortality among the common malignancies and development of novel therapies for PDAC through identification of good molecular targets is an urgent issue. Among dozens of over-expressing genes identified through our gene-expression profile analysis of PDAC cells, we here report CST6 (Cystatin 6 or E/M) as a candidate of molecular targets for PDAC treatment . Reverse transcriptase-polymerase chain reaction (RT-PCR) and immunohistochemical analysis confirmed overexpression of CST6 in PDAC cells, but no or limited expression of CST6 was observed in normal pancreas and other vital organs. Knockdown of endogenous CST6 expression by small interfering RNA attenuated PDAC cell growth, suggesting its essential role in maintaining viability of PDAC cells. Concordantly, constitutive expression of CST6 in CST6-null cells promoted their growth *in vitro* and *in vivo*. Furthermore, the addition of mature recombinant CST6 in culture medium also promoted cell proliferation in a dose-dependent manner, whereas recombinant CST6 lacking its proteinase-inhibitor domain and its non-glycosylated form did not. Overexpression of CST6 inhibited the intracellular activity of cathepsin B, which is one of the putative substrates of CST6 proteinase inhibitor and can intracellularly function as a pro-apoptotic factor. These findings imply that CST6 is likely to involve in the proliferation and survival of pancreatic cancer probably through its proteinase inhibitory activity, and it is a promising molecular target for development of new therapeutic strategies for PDAC.

C2orf18 (ANTBP)

Through our genome-wide gene expression profiles of microdissected PDAC cells, we here identified a novel gene *C2orf18* as a molecular target for PDAC treatment. Transcriptional and immunohistochemical analysis validated its overexpression in PDAC cells and limited expression in normal adult organs. Knockdown of *C2orf18* by small-interfering RNA in PDAC cell lines resulted in induction of apoptosis and suppression of cancer cell growth, suggesting its essential role in maintaining viability of PDAC cells. We showed that *C2orf18* was localized in the mitochondria and it could interact with adenine nucleotide translocase 2 (ANT2), which is involved in maintenance of the mitochondrial membrane potential and energy homeostasis, and was indicated some roles in apoptosis. These findings implicated that C2orf18, termed ANT2-binding protein (ANT2BP), might serve as a candidate molecular target for pancreatic cancer therapy.

(6) Prostate cancer

STC2 (stanniocalcin 2)

Prostate cancer is usually androgen-dependent and responds well to androgen ablation therapy based on castration. However, at a certain stage some prostate cancers eventually acquire a castration-resistant phenotype where they progress aggressively and show very poor response to any anticancer therapies. To characterize the molecular features of these clinical castration-resistant prostate cancers, we previously analyzed gene expression profiles by genome-wide cDNA microarrays combined with microdissection and found dozens of *trans*-activated genes in clinical castration-resistant prostate cancers. Among them, we report the identification of a new biomarker, stanniocalcin 2 (STC2), as an overexpressed gene in castration-resistant prostate cancer cells. Real-time polymerase chain reaction and immunohistochemical analysis confirmed overexpression of STC2, a 302-amino-acid glycoprotein hormone, specifically in castrationresistant prostate cancer cells and aggressive castration-naïve prostate cancers with high Gleason scores (8-10). The gene was not expressed in normal prostate, nor in most indolent castration-naïve prostate cancers. Knockdown of STC2 expression by short interfering RNA in a

prostate cancer cell line resulted in drastic attenuation of prostate cancer cell growth. Concordantly, STC2 overexpression in a prostate cancer cell line promoted prostate cancer cell growth, indicating its oncogenic property. These findings suggest that STC2 could be involved in aggressive phenotyping of prostate cancers, including castration-resistant prostate cancers, and that it should be a potential molecular target for development of new therapeutics and a diagnostic biomarker for aggressive prostate cancers.

(7) Thyroid cancer

In order to clarify the molecular mechanism involved in thyroid carcinogenesis and to identify candidate molecular targets for diagnosis and treatment, we analyzed genome-wide gene expression profiles of 18 papillary thyroid carcinomas with a microarray representing 38,500 genes in combination with laser microbeam microdissection. We identified 243 transcripts that were commonly up-regulated and 138 transcripts that were down-regulated in thyroid carcinoma. Among these 243 transcripts identified, only 71 transcripts were reported as up-regulated genes in previous microarray studies, in which bulk cancer tissues and normal thyroid tissues were used for the analysis. We further selected genes that were overexpressed very commonly in thyroid carcinoma, though were not expressed in the normal human tissues examined. Among them, we focused on the regulator of G-protein signaling 4 (*RGS4*) and knocked-down its expression in thyroid cancer cells by small-interfering RNA. The effective down-regulation of its expression levels in thyroid cancer cells significantly attenuated viability of thyroid cancer cells, indicating the significant role of RGS4 in thyroid carcinogenesis. Our data should be helpful for a better understanding of the tumorigenesis of thyroid cancer and could contribute to the development of diagnostic tumor markers and molecular-targeting therapy for patients with thyroid cancer.

(8) Ovarian cancer

We aimed to clarify the molecular mechanisms involved in ovarian carcinogenesis, and to identify candidate molecular targets for its diagnosis and treatment. The genome-wide gene expression profiles of 22 epithelial ovarian carcinomas were analyzed with a microarray representing 38,500 genes, in combination with laser microbeam microdissection. A total of 273 commonly up-regulated transcripts and 387 down-regulated transcripts were identified in the ovarian carcinoma samples. Of the 273 up-regulated transcripts, only 87 (31.9%) were previously reported as upregulated in microarray studies using bulk cancer tissues and normal ovarian tissues for analysis. *CHMP4C* (chromatinmodifying protein 4C) was frequently overexpressed in ovarian carcinoma tissue, but not expressed in the normal human tissues used as a control. Our data should contribute to an improved understanding of tumorigenesis in ovarian cancer, and aid in the development of diagnostic tumor markers and molecular-targeting therapy for patients with the disease.

(9) Proteomics

To screen for glycoproteins showing aberrant sialylation patterns in sera of cancer patients and apply such information for biomarker identification, we performed SELDI-TOF MS analysis coupled with lectin-coupled ProteinChip arrays (Jacalin or SNA) using sera obtained from lung cancer patients and control individuals. Our approach consisted of three processes, (1) removal of 14 abundant proteins in serum, (2) enrichment of glycoproteins with lectin-coupled ProteinChip arrays, and (3) SELDI-TOF MS analysis with acidic glycoprotein-compatible matrix. We identified 41 protein peaks showing significant differences ($P < 0.05$) in the peak levels between the cancer and control groups using the Jacalin- and SNA- ProteinChips. Among them, we identified loss of Neu5Ac (α2, 6) Gal/GalNAc structure in apolipoprotein C-III (apoC-III) in cancer patients through subsequent MALDI-QIT-TOF MS/MS. Furthermore, subsequent validation experiments using an additional set of 60 lung adenocarcinoma patients and 30 normal controls demonstrated that there is a higher frequency of serum apoC-III with loss of α2, 6-linkage Neu5Ac residues in lung cancer patients compared to controls. Our results have demonstrated that lectin-coupled ProteinChip technology allows the high-throughput and specific recognition of cancer-associated aberrant glycosylations, and implied a possibility of its applicability to studies on other diseases.

(10) Chemosensitivity

Breast Cancer

Neoadjuvant chemotherapy with docetaxel for advanced breast cancer can improve the radicality for a subset of patients, but some patients suffer from severe adverse drug reactions without any benefit. To establish a method for predicting responses to docetaxel, we analyzed gene expression profiles of biopsy materials from 29 advanced breast cancers using a cDNA

microarray consisting of 36,864 genes or ESTs, after enrichment of cancer cell population by laser microbeam microdissection. Analyzing eight PR (partial response) patients and twelve patients with SD (stable disease) or PD (progressive disease) response, we identified dozens of genes that were expressed differently between the 'responder (PR)' and 'non-responder (SD or PD)' groups. We further selected the nine 'predictive' genes showing the most significant differences and established a numerical prediction scoring system that clearly separated the responder group from the non-responder group. This system accurately predicted the drug responses of all of nine additional test cases that were reserved from the original 29 cases. Moreover, we developed a quantitative PCR-based prediction system that could be feasible for routine clinical use. Our results suggest that the sensitivity of an advanced breast cancer to the neoadjuvant chemotherapy with docetaxel could be predicted by expression patterns in this set of genes.

2. Pharmacogenomics

(1) Warfarin maintenance-dose requirements

The International Warfarin Pharmacogenetics Consortium

Genetic variability among patients plays an important role in determining the dose of warfarin that should be used when oral anticoagulation is initiated, but practical methods of using genetic information have not been evaluated in a diverse and large population. We developed and used an algorithm for estimating the appropriate warfarin dose that is based on both clinical and genetic data from a broad population base. Clinical and genetic data from 4043 patients were used to create a dose algorithm that was based on clinical variables only and an algorithm in which genetic information was added to the clinical variables. In a validation cohort of 1009 subjects, we evaluated the potential clinical value of each algorithm by calculating the percentage of patients whose predicted dose of warfarin was within 20% of the actual stable therapeutic dose; we also evaluated other clinically relevant indicators. In the validation cohort, the pharmacogenetic algorithm accurately identified larger proportions of patients who required 21 mg of warfarin or less per week and of those who required 49 mg or more per week to achieve the target international normalized ratio than did the clinical algorithm (49.4% vs. 33.3%, $P<0.001$, among patients requiring$<$or$=$21 mg per week; and 24.8% vs. 7.2%, $P<0.001$, among those requiring$>$or$=$49 mg per week). The use of a pharmacogenetic algorithm for estimating the appropriate initial dose of warfarin produces recommendations that are significantly closer to the required stable therapeutic dose than those derived from a clinical algorithm or a fixed-dose approach. The greatest benefits were observed in the 46.2% of the population that required 21 mg or less of warfarin per week or 49 mg or more per week for therapeutic anticoagulation.

(2) Genotype of *CYP2D6* and selection of adjuvant hormonal therapy with tamoxifen for breast cancer patients

Authors: Kazuma Kiyotani[1], Taisei Mushiroda[1], Mitsunori Sasa[2], Yoshimi Bando[3], Ikuko Sumitomo[2], Naoya Hosono[4], Michiaki Kubo[4], Yusuke Nakamura[1,5], and Hitoshi Zembutsu[5]: [1]Laboratory for Pharmacogenetics, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN), [2]Department of Surgery, Tokushima Breast Care Clinic, [3]Department of Molecular and Environmental Pathology, Institute of Health Biosciences, The University of Tokushima Graduate School, [4]Laboratory for genotyping, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN), [5]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, The University of Tokyo

The clinical outcomes of breast cancer patients treated with tamoxifen may be influenced by the activity of cytochrome P450 2D6 (CYP2D6) enzyme because tamixifen is metabolized by CYP2D6 to its active forms of antiestrogenic metabolite, 4-hydroxytamoxifen and endoxifen. We investigated the predictive value of the *CYP2D6**10 allele, which decreased CYP2D6 activity, for clinical outcomes of patients that received adjuvant tamoxifen monotherapy after surgical operation on breast cancer. Among 67 patients examined, those homozygous for the *CYP2D6**10 alleles revealed a significantly higher incidence of recurrence within 10 years after the operation ($P=0.0057$; odds ratio, 16.63; 95% confidence interval, 1.75-158.12), compared with those homozygous for the wild-type *CYP2D6**1 alleles. The elevated risk of recurrence seemed to be dependent on the number of *CYP2D6**10 alleles ($P=0.0031$ for trend). Cox proportional hazard analysis demonstrated that the *CYP2D6* genotype and tumor size were independent factors affecting recurrence-free survival. Patients with the *CYP2D6**10/*10 genotype showed a significantly shorter recurrence-free survival period ($P=0.036$; adjusted hazard

ratio, 10.04; 95% confidence interval, 1.17-86.27) compared to patients with *CYP2D6**1/*1 after adjustment of other prognosis factors. The present study suggests that the CYP2D6 genotype should be considered when selecting adjuvant hormonal therapy for breast cancer patients.

(3) Genotype of drug metabolism/transporter genes and Docetaxel-induced leukopenia/neutropenia

Authors: Kazuma Kiyotani[1], Taisei Mushiroda[1], Michiaki Kubo[2], Hitoshi Zembutsu[3], Yuichi Sugiyama[4], and Yusuke Nakamura[1,3]: [1]Laboratory for Pharmacogenetics, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN), [2]Laboratory for genotyping, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN), [3]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, The University of Tokyo, [4]Department of Molecular Pharmacokinetics, Graduate School of Pharmaceutical Sciences, The University of Tokyo

Despite long-term clinical experience with docetaxel, unpredictable severe adverse reactions remain an important determinant for limiting the use of the drug. To identify a genetic factor (s) determining the risk of docetaxel-induced leukopenia/neutropenia, we selected subjects who received docetaxel chemotherapy from samples recruited at BioBank Japan, and conducted a case-control association study. We genotyped 84 patients, 28 patients with grade 3 or 4 leukopenia/neutropenia, and 56 with no toxicity (patients with grade 1 or 2 were excluded), for a total of 79 single nucleotide polymorphisms (SNPs) in seven genes possibly involved in the metabolism or transport of this drug: *CYP3A4, CYP3A5, ABCB1, ABCC2, SLCO1B3, NR1I2*, and *NR1I3*. Since one SNP in *ABCB1*, four SNPs in *ABCC2*, four SNPs in *SLCO1B3*, and one SNP in *NR1I2* showed a possible association with the grade 3 leukopenia/neutropenia (P-value of<0.05), we further examined these 10 SNPs using 29 additionally obtained patients, 11 patients with grade 3/4 leukopenia/neutropenia, and 18 with no toxicity. The combined analysis indicated a significant association of rs 12762549 in *ABCC2* ($P=0.00022$) and rs11045585 in *SLCO1B3* ($P=0.00017$) with docetaxel-induced leukopenia/neutropenia. When patients were classified into three groups by the scoring system based on the genotypes of these two SNPs, patients with a score of 1 or 2 were shown to have a significantly higher risk of docetaxel-induced leukopenia/neutropenia as compared to those with a score of 0 ($P=0.0000057$; odds ratio [OR], 7.00; 95% CI [confidence interval], 2.95-16.59). This prediction system correctly classified 69.2% of severe leukopenia / neutropenia and 75.7% of non-leukopenia/neutropenia into the respective categories, indicating that SNPs in *ABCC2* and *SLCO1B3* may predict the risk of leukopenia/neutropenia induced by docetaxel chemotherapy.

(4) HLA genotype and Nevirapine (NVP)-induced skin rash

Authors: Soranun Chantarangsu[1,2], Taisei Mushiroda[1], Surakameth Mahasirimongkol[5], Sasisopin Kiertiburanakul[3], Somnuek Sungkanuparph[3], Weerawat Manosuthi[6], Woraphot Tantisiriwat[7], Angkana Charoenyingwattana[4], Thanyachai Sura[3], Wasun Chantratita[2], and Yusuke Nakamura[1]: [1]Research Group for Pharmacogenomics, RIKEN, Center for Genomic Medicine Departments of [2]Pathology, [3]Medicine, Faculty of Medicine, [4]Department of Pharmacy, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand, [5]Center for International Cooperation, Department of Medical Sciences, [6]Bamrasnaradura Infectious Diseases Institute, Ministry of Public Health, [7]Department of Preventive Medicine, Faculty of Medicine, Srinakharinwirot University, Nakornnayok, Thailand

We investigated a possible involvement of differences in human leukocyte antigens (HLA) in the risk of nevirapine (NVP)-induced skin rash among HIV-infected patients by a step-wise case-control association study. We first genotyped by a sequence-based HLA typing method for the *HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQB1*, and *HLA-DPB1* in the first set of samples consisted of 80 samples from patients with NVP-induced skin rash and 80 samples from NVP-tolerant patients. Subsequently, we verified HLA alleles that showed a possible association in the first screening using an additional set of samples consisting of 67 cases with NVP-induced skin rash and 105 controls. An *HLA-B**3505 allele revealed a significant association with NVP-induced skin rash in the first and second screenings. In the combined data set, the *HLA-B**3505 allele was observed in 17.5% of the patients with NVP-induced skin rash compared with only 1.1% observed in NVP-tolerant patients [odds ratio (OR)=18.96; 95% confidence interval (CI)=4.87-73.44, Pc=4.6×10] and 0.7% in general Thai population (OR=29.87; 95% CI =5.04-175.86, Pc=2.6×10). The logistic regression analysis also indicated *HLA-B**3505 to be

significantly associated with skin rash with OR of 49.15 (95% CI=6.45-374.41, $P=0.00017$). We suggest that strong association between the *HLA-B**3505 and NVP-induced skin rash provides a novel insight into the pathogenesis of drug-induced rash in the HIV-infected population. On account of its high specificity (98.9%) in identifying NVP-induced rash, it is possible to utilize the *HLA-B**3505 as a marker to avoid a subset of NVP-induced rash, at least in Thai population.

3. Common diseases

(1) Chronic hepatitis B

Authors: Yoichiro Kamatani[1,2], Sukanya Wattanapokayakit[3], Hidenori Ochi[4,5], Takahisa Kawaguchi[4], Atsushi Takahashi[4], Naoya Hosono[4], Michiaki Kubo[4], Tatsuhiko Tsunoda[4], Naoyuki Kamatani[4], Hiromitsu Kumada[6], Aekkachai Puseenam[7], Thanyachai Sura[7], Yataro Daigo[2], Kazuaki Chayama[4,5], Wasun Chantratita[8], Yusuke Nakamura[1,4], and Koichi Matsuda[1]: [1]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, The University of Tokyo, [2]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The Universtiy of Tokyo, [3]Center for International Cooperation, Department of Medical Sciences, Ministry of Public Health, Thailand, [4]Center for Genomic Medicine, RIKEN, [5]Department of Medicine and Molecular Science, Division of Frontier Medical Science, Programs for Biomedical Research, Graduate School of Biomedical Sciences, Hiroshima University, [6]Department of Hepatology, Toranomon Hospital, [7]Department of Medicine, Faculty of Medicine and [8]Virology and Molecular Microbiology Unit, Department of Pathology, Faculty of Medicine, Ramathidi Hospital, Mahidol University, Thailand

Chronic hepatitis B is a serious infectious liver disease that often progresses to liver cirrhosis and hepatocellular carcinoma; however, clinical outcomes after viral exposure enormously vary among individuals. Through a two-step genome-wide association study using 786 Japanese chronic hepatitis B patients and 2,201 controls, here we identified a significant association of chronic hepatitis B with 11 SNPs in a region including *HLA-DPA1* and *HLA-DPB1* genes. These associations were validated in two Japanese and one Thai cohorts consisting of 1,300 cases and 2,100 controls (combined $P=6.34\times10^{-39}$ and 2.31×10^{-38}, OR=0.57 and 0.56, respectively). Subsequent analyses revealed disease susceptible haplotypes (*HLA-DPA1**0202-*DPB1**0501 and *HLA-DPA1**0202-*DPB1**0301, OR=1.45 and 2.31, respectively) and protective haplotypes (*HLA-DPA1**0103-*DPB1**0402 and *HLA-DPA1**0103-*DPB1**0401, OR=0.52 and 0.57, respectively). Our findings demonstrated that genetic variations in the *HLA-DP* locus are strongly associated with the risk of persistent infection of hepatitis B virus.

(2) Idiopathic pulmonary fibrosis (IPF)

Authors: Taisei Mushiroda[1], Sukanya Wattanapokayakit[2], Atsushi Takahashi[3], Toshihiro Nukiwa[4], Shoji Kudoh[5], Takashi Ogura[6], Hiroyuki Taniguchi[7], Michiaki Kubo[8], Naoyuki Kamatani[3], Yusuke Nakamura[1,9], and the Pirfenidone Clinical Study Group[4]: [1]Laboratory for Pharmacogenetics, Institute of Physical and Chemical Research (RIKEN), [2]Laboratory for Cardiovascular Diseases, Institute of Physical and Chemical Research (RIKEN), [3]Laboratory of Statistical Analysis, Institute of Physical and Chemical Research (RIKEN), [4]Department of Respiratory Oncology and Molecular Medicine, Institute of Development, Aging, and Cancer, Tohoku University, [5]Fourth Department of Internal Medicine, Nippon Medical School, [6]Department of Respiratory Medicine, Kanagawa Cardiovascular and Respiratory Center, [7]Department of Respiratory Medicine and Allergy, Tosei General Hospital, Aichi, [8]Laboratory for genotyping, Institute of Physical and Chemical Research (RIKEN), [9]Laboratory of Molecular Medicine, Institute of Medical Science, University of Tokyo

In order to identify a gene (s) susceptible to idiopathic pulmonary fibrosis (IPF), we conducted a genome-wide association (GWA) study by genotyping 159 patients with IPF and 934 controls for 214,508 tag single-nucleotide polymorphisms (SNPs). We further evaluated selected SNPs in a replication sample set (83 cases and 535 controls) and found a significant association of an SNP in intron 2 of the *TERT* gene (rs2736100), which encodes a reverse transcriptase that is a component of a telomerase, with IPF; a combination of two data sets revealed a p value of 2.9×10 (-8) (GWA, 2.8×10 (-6); replication, 3.6×10 (-3)). Considering previous reports indicating that rare mutations of *TERT* are found in patients with familial IPF, we suggest that the common genetic variation within *TERT* may contribute to the risk of sporadic IFP in the Japanese population.

(3) Schizophrenia

Authors: Elitza T Betcheva[1], Taisei Mushi-

roda[2], Atsushi Takahashi[3], Michiaki Kubo[4], Sena K Karachanak[5], Irina T Zaharieva[6], Radoslava V Vazharova[5], Ivanka I Dimova[5], Vihra K Milanova[6], Todor Tolev[7], George Kirov[8], Michael J Owen[8], Michael C O'Donovan[8], Naoyuki Kamatani[3], Yusuke Nakamura[9], and Draga I Toncheva[5]: [1]Laboratory for Cardiovascular Diseases, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN), [2]Laboratory for Pharmacogenetics, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN), [3]Laboratory of Statistical Analysis, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN), [4]Laboratory for Genotyping, SNP Research Center, The Institute of Physical and Chemical Research (RIKEN), [5]Department of Medical Genetics, Medical Faculty, Medical University, Sofia, Bulgaria, [6]Department of Psychiatry, Aleksandrovska Hospital, Medical University, Sofia, Bulgaria, [7]Department of Psychiatry, Dr Georgi Kisiov Hospital, Radnevo, Bulgaria, [8]Department of Psychological Medicine, Cardiff University, School of Medicine, Henry Wellcome Building, Heath Park, Cardiff, UK, [9]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, The University of Tokyo

The development of molecular psychiatry in the last few decades identified a number of candidate genes that could be associated with schizophrenia. A great number of studies often result with controversial and non-conclusive outputs. However, it was determined that each of the implicated candidates would independently have a minor effect on the susceptibility to that disease. Herein we report results from our replication study for association using 255 Bulgarian patients with schizophrenia and schizoaffective disorder and 556 Bulgarian healthy controls. We have selected from the literatures 202 single nucleotide polymorphisms (SNPs) in 59 candidate genes, which previously were implicated in disease susceptibility, and we have genotyped them. Of the 183 SNPs successfully genotyped, only 1 SNP, rs6277 (C957T) in the DRD2 gene ($P = 0.0010$, odds ratio = 1.76), was considered to be significantly associated with schizophrenia after the replication study using independent sample sets. Our findings support one of the most widely considered hypotheses for schizophrenia etiology, the dopaminergic hypothesis.

Publications

1. 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-RETINA): a novel SNP-based method for detecting allelic asymmetries within copy number variation regions. Hum. Mutation 29: 182-189, 2008.
2. 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.
3. 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. 27: 2686-2692, 2008.
4. 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. 14: 1333-1339, 2008
5. M. Kato, F. Miya, Y. Kanemura, T. Tanaka, Y. Nakamura, and T. Tsunoda: Recombination rates of genes expressed in human tissues. Hum. Mol. Genet. 17: 577-586, 2008.
6. Leung, A.A.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.
7. 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.
8. 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.
9. 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.
10. 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. 99: 432-440, 2008.
11. 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. 10: 908-920, 2008.
12. 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. 17: 617-627, 2008.
13. 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.
14. Daigo, Y. and Nakamura Y. From cancer genomics to thoracic oncology: discovery of new biomarkers and therapeutic targets for lung and esophageal carcinoma. (Review Article) General Thoracic and Cardiovascular Surgery 56: 43-53, 2008.
15. Kiyotani, K., Mushiroda, T., Kubo, M., Zembutsu, H., Sugiyama, Y. and Nakamura, Y. Association of genetic polymorphisms in SLCO1B3 and ABCC2 with docetaxel-induced leukopenia. Cancer Sci. 99: 967-972, 2008.
16. Kiyotani, K., Mushiroda, T., Sasa, M., Bando, Y., Sumitomo, I., Hosono, N., Kubo, M., Nakamura, Y. and Zembutsu, H. Impact of *CYP2D6**10 on recurrence-free survival in breast cancer patients receiving adjuvant tamoxifen therapy. Cancer Sci. 99: 995-999, 2008.
17. Kato, T., Sato, N., Takano, A., Miyamoto, M., Nishimura, H., Tsuchiya, E., Kondo, S., Nakamura, Y. and Daigo, Y. Activation of Placenta-Specific Transcription Factor Distalless Homeobox 5 Predicts Clinical Outcome in Primary Lung Cancer Patients. Clin. Cancer Res. 14: 2363-2370, 2008.
18. Tenesa, A., Farrington, S.M., Prendergast, J. G., Porteous, M.E., Walker, M., Haq, N., Barnetson, R.A., Theodoratou, E., Cetnarskyj, R., Cartwright, N., Semple, C., Clark, A.J., Reid, F.J., Smith, L.A., Kavoussanakis, K., Koessler, T., Pharoah, P.D., Buch, S., Schafmayer, C., Tepel, J., Schreiber, S., Völzke, H., Schmidt, C.O., Hampe, J., Chang-Claude, J., Hoffmeister, M., Brenner, H., Wilkening, S., Canzian, F., Capella, G., Moreno, V., Deary, I.J., Starr, J.M., Tomlinson, I.P., Kemp, Z., Howarth, K., Carvajal-Carmona, L., Webb, E., Broderick, P., Vijayakrishnan, J., Houlston, R.S., Rennert, G., Ballinger, D., Rozek, L., Gruber, S.B., Matsuda, K., Kidokoro, T., Nakamura, Y., Zanke, B.W., Greenwood, C. M., Rangrej, J., Kustra, R., Montpetit, A., Hudson, T.J., Gallinger, S., Campbell, H. and Dunlop, M.G. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q 24 and 18q21. Nat. Genet. 40: 631-637, 2008.
19. Mototani H., Iida, A., Nakajima, M., Furuichi, T., Miyamoto, Y., Tsunoda, T., Sudo, A., Kotani, A., Uchida, K., Ozaki, K., Tanaka, Y., Nakamura, Y., Tanaka, T., Notoya, K. and Ikegawa, S.A functional SNP in EDG2 increases susceptibility to knee osteoarthritis in Japanese. Hum. Mol. Genet. 17: 1790-1797, 2008.
20. Mizukami, Y., Kono, K., Daigo, Y., Takano, A., Tsunoda, T., Kawaguchi, Y., Nakamura, Y. and Fujii, H. Detection of novel Cancer-Testis antigen-specific T-cell responses in TIL, regional lymph nodes and PBL in patients with esophageal squamous cell carcinoma. Cancer Sci. 99: 1448-1454, 2008.
21. Mushiroda, T., Wattanapokayakit, S., Takahashi, A., Nukiwa, T., Kudoh, S., Ogura, T., Taniguchi, H., Pirfenidone Clinical Study Group., Kubo, M., Kamatani, N. and Nakamura Y.A genome-wide association study identifies an association of a common variant in TERT with susceptibility to idiopathic pulmonary fibrosis. J. Med. Genet. 45: 654-656, 2008.
22. Hosokawa, M., Kashiwaya, K., Furihara, M.,

Eguchi, H., Ohigashi, H., Ishikawa, O., Shinomura, Y., Imai, K., Nakamura, Y. and Nakagawa H. Overexpression of cysteine proteinase inhibitor cystatin 6 promotes pancreatic cancer growth. Cancer Sci. 99: 1626-1632, 2008.

23. Study Group of Millennium Genome Project for Cancer, Sakamoto, H., Yoshimura ,K., Saeki, N., Katai, H., Shimoda, T., Matsuno, Y., Saito, D., Sugimura, H., Tanioka, F., Kato, S., Matsukura, N., Matsuda, N., Nakamura, T., Hyodo, I., Nishina, T., Yasui, W., Hirose, H., Hayashi, M., Toshiro, E., Ohnami, S., Sekine, A., Sato, Y., Totsuka, H., Ando, M., Takemura, R., Takahashi, Y., Ohdaira, M., Aoki, K., Honmyo, I., Chiku, S., Aoyagi, K., Sasaki, H., Ohnami, S., Yanagihara, K., Yoon, K.A., Kook, M.C., Lee, Y.S., Park, S.R., Kim, C.G., Choi, I.J., Yoshida, T., Nakamura, Y. and Hirohashi, S. Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. Nat. Genet. 40: 730-740, 2008.
24. Ueki, T., Nishidate, T., Park, J.H., Lin, M.L., Shimo, A., Hirata, K., Nakamura, Y. and Katagiri, T. Involvement of elevated expression of multiple cell-cycle regulator, DTL/RAMP (denticleless/RA-regulated nuclear matrix associated protein), in the growth of breast cancer cells. Oncogene. 27: 5672-5683, 2008.
25. Miyamoto, Y., Shi, D., Nakajima, M., Ozaki, K., Sudo, A., Kotani, A., Uchida, A., Tanaka, T., Fukui, N., Tsunoda, T., Takahashi, A., Nakamura, Y., Jiang, Q. and Ikegawa S. Common variants in DVWA on chromosome 3p24.3 are associated with susceptibility to knee osteoarthritis. Nat. Genet. 40: 994 -998, 2008.
26. Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G., Ng, D.P., Holmkvist, J., Borch-Johnsen, K., Jørgensen, T., Sandbaek, A., Lauritzen, T., Hansen, T., Nurbaya, S., Tsunoda, T., Kubo, M., Babazono, T., Hirose, H., Hayashi, M., Iwamoto, Y., Kashiwagi, A., Kaku, K., Kawamori, R., Tai, E.S., Pedersen, O., Kamatani, N., Kadowaki, T., Kikkawa, R., Nakamura, Y. and Maeda, S. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat. Genet. 40: 1098-1102, 2008.
27. Harao, M., Hirata, S., Irie, A., Senju, S., Nakatsura, T., Komori, H., Ikuta, Y., Yokomine, K., Imai, K., Inoue, M., Harada, K., Mori, T., Tsunoda, T., Nakatsuru, S., Daigo, Y., Nomori, H., Nakamura, Y., Baba, H. and Nishimura, Y. HLA-A2-restricted CTL epitopes of a novel lung cancer-associated cancer testis antigen, cell division cycle associated 1, can induce tumor-reactive CTL. Int. J. Cancer 123: 2616-2625, 2008.
28. Imai, K., Hirata, S., Irie, A., Senju, S., Ikuta, Y., Yokomine, K., Harao, M., Inoue, M., Tsunoda, T., Nakatsuru, S., Nakagawa, H., Nakamura, Y., Baba, H. and Nishimura, Y. Identification of a novel tumor-associated antigen, cadherin 3/P-cadherin, as a possible target for immunotherapy of pancreatic, gastric and colorectal cancers. Clin. Cancer Res. 14: 6487-6495, 2008.
29. Nikolova, D.N., Zembutsu, H., Sechanov, T., Vidinov, K., Kee, L.S., Ivanova, R., Becheva, E., Kocova, M., Toncheva, D. and Nakamura, Y. Identification of molecular targets for treatment of thyroid carcinoma. Oncol. Rep. 20: 105-121, 2008.
30. Nakamura Y. Pharmacogenomics and drug toxicity. (Editorial) New. Eng. J. Med. 359: 856-858, 2008.
31. Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y. and Shirakawa, M. Hemimethylated DNA recognition by the SRA protein Np95 via a base flipping mechanism. Nature 455: 818-821, 2008.
32. Inoue, H., Iga, M., Nabeta, H., Yokoo, T., Suehiro, Y., Okano, S., Inoue, M., Kinoh, H., Katagiri, T., Takayama, K., Yonemitsu, Y., Hasegawa, M., Nakamura, Y., Nakanishi, Y. and Tani, K. Non-transmissible SeV encoding GM-CSF is a novel and potent vector system to produce autologous tumor vaccines. Cancer Sci. 99: 2315-2326, 2008.
33. Konda R, Sugimura J, Sohma F, Katagiri T, Nakamura Y, Fujioka T. Over expression of hypoxia-inducible protein 2, hypoxia-inducible factor-1αand nuclear factor κB is putatively involved in acquired renal cyst formation and subsequent tumor transformation in patients with end stage renal failure. J. Urol. 180: 481-485, 2008.
34. Hotta, K., Nakata, Y., Matsuo, T., Kamohara, S., Kotani, K., Komatsu, R., Itoh, N., Mineo, I., Wada, J., Masuzaki, H., Yoneda, M., Nakajima, A., Miyazaki, S., Tokunaga, K., Kawamoto, M., Funahashi, T., Hamaguchi., K., Yamada, K., Hanafusa, T., Oikawa, S., Yoshimatsu, H., Nakao, K., Sakata, T., Matsuzawa, Y., Tanaka, K., Kamatani, N. and Nakamura Y. Variations in the FTO gene are associated with severe obesity in the Japanese. J. Hum. Genet. 53: 546-553, 2008.
35. Kato, M., Nakamura, Y. and Tsunoda, T. An algorithm for inferring complex haplotypes in a region of copy-number variation. Am. J. Hum. Genet. 83: 157-169, 2008.
36. Kato, M., Nakamura, Y. and Tsunoda, T. MOCSphaser: a haplotype inference tool

from a mixture of copy number variation and single nucleotide polymorphism data. Bioinformatics. 24: 1645-1646, 2008.
37. Yasuda, K., Miyake, K., Horikawa, Y., Hara, K., Osawa, H., Furuta, H., Hirota, Y., Mori, H., Jonsson, A., Sato, Y., Yamagata, K., Hinokio, Y., Wang, H.Y., Tanahashi, T., Nakamura, N., Oka, Y., Iwasaki, N., Iwamoto, Y., Yamada, Y., Seino, Y., Maegawa, H., Kashiwagi, A., Takeda, J., Maeda, E., Shin, H.D., Cho, Y.M., Park, K.S., Lee, H.K., Ng, M.C., Ma, R.C., So, W.Y., Chan, J.C., Lyssenko, V., Tuomi, T., Nilsson, P., Groop, L., Kamatani, N., Sekine, A., Nakamura, Y., Yamamoto, K., Yoshida, T., Tokunaga, K., Itakura, M., Makino, H., Nanjo, K., Kadowaki, T. and Kasuga M. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. Nat. Genet. 40: 1092-1097, 2008.
38. Yamaguchi-Kabata, Y., Nakazono, K., Takahashi, A., Saito, S., Hosono, N., Kubo, M., Nakamura, Y. and Kamatani N. Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: Effects on population-based association studies. Am. J. Hum. Genet. 83: 445-456, 2008.
39. Okada, Y., Mori, M., Yamada, R., Suzuki, A., Kobayashi, K., Kubo, M., Nakamura, Y. and Yamamoto, K. SLC22A4 polymorphism and rheumatoid arthritis susceptibility: A replication study in a Japanese population and a metaanalysis. J. Rheumatol. 35: 1723-1728, 2008.
40. Omori, S., Tanaka, Y., Takahashi, A., Hirose, H., Kashiwagi, A., Kaku, K., Kawamori, R., Nakamura, Y. and Maeda, S. Association of CDKAL1, IGF2BP2, CDKN2A/B, HHEX, SLC30A8, and KCNJ11 with susceptibility of type 2 diabetes in a Japanese population. Diabetes. 57: 791-795, 2008.
41. Misawa, K., Fujii, S., Yamazaki, T., Takahashi, A., Takasaki, J., Yanagisawa, M., Ohnishi, Y., Nakamura, Y. and Kamatani, N. New correction algorithms for multiple comparisons in case-control multilocus association studies based on haplotypes and diplotype configurations. J. Hum. Genet. 53: 789-801, 2008.
42. Chantarangsu, S., Mushiroda, T., Mahasirimongkol, S., Kiertiburanakul, S., Sungkanuparph, S., Manosuthi, W., Tantisiriwat, W., Charoenyingwattana, A., Sura, T., Chantratita, W. and Nakamura, Y. *HLA-B*3505* allele is a strong predictor for nevirapine-induced skin adverse drug reactions in Thai HIV-infected patients . Pharmacogenet . Genomics. 19: 139-146, 2009.
43. Suzuki, A., Yamada, R., Kochi, Y., Sawada, T., Okada, Y., Matsuda, K., Kamatani, Y., Mori, M., Shimane, K., Hirabayashi, Y., Takahashi, A., Tsunoda, T., Miyatake, A., Kubo, M., Kamatani, N., Nakamura, Y. and Yamamoto K. Functional SNPs in CD244 increase the risk of rheumatoid arthritis in a Japanese population. Nat. Genet. 40: 1224-1229, 2008.
44. Yamazaki, K., Takahashi, A., Takazoe, M., Kubo, M., Onouchi, Y., Fujino, A., Kamatani, N., Nakamura, Y. and Hata A. Positive association of genetic variants in the upstream region of NXT2-3 with Crohn's disease in Japanese patients. Gut. 58: 228-232, 2009.
45. Nikolova, D.N., Doganov, N., Dimitrov, R., Angelov, K., Kee, L.S., Dimova, I., Toncheva, D., Nakamura, Y. and Zembutsu H. Genome-wide gene expression profiles of ovarian carcinoma: identification of molecular targets for treatment of ovarian carcinoma. Mol. Med. Rep., in press, 2008.
46. Hotta, K., Nakamura, M., Nakata, Y., Matsuo, T., Kamohara, S., Kotani, K., Komatsu, R., Itoh, N., Mineo, I., Wada, J., Masuzaki, H., Yoneda, M., Nakajima, A., Miyazaki, S., Tokunaga, K., Kawamoto, M., Funahashi, T., Hamaguchi, K., Yamada, K., Hanafusa, T., Oikawa, S., Yoshimatsu, H., Nakao, K., Sakata, T., Matsuzawa, Y., Tanaka, K., Kamatani, N. and Nakamura Y. INSIG2 gene rs7566605 polymorphism is associated with severe obesity in Japanese. J. Hum. Genet. 53: 857-862, 2008.
47. Iwahori, K., Osaki, T., Serada, S., Fujimoto, M., Suzuki, H., Kishi, Y., Yokoyama, A., Hamada, H., Fujii, Y., Yamaguchi, K., Hirashima, T., Matsui, K., Tachibana, I., Nakamura, Y., Kawase, I. and Naka, T. Megakaryocyte potentiating factor as a tumor maker of malignant pleural mesothelioma: Evaluation in comparison with mesothelin. Lung Cancer. 62: 45-54, 2008.
48. Hirota, T., Harada, M., Sakashita, M., Doi, S., Miyatake, A., Fujita, K., Enomoto, T., Ebisawa, M., Yoshihara, S., Noguchi, E., Saito, H., Nakamura, Y. and Tamari, M. Genetic polymorphism regulating ORM1-like 3 (Saccharomyces cerevisiae) expression is associated with childhood atopic asthma in a Japanese population. J. Allergy Clin. Immunol. 121: 769-770, 2008
49. Harada, M., Hirota, T., Jodo, A.I., Doi, S., Kameda, M., Fujita, K., Miyatake, A., Enomoto, T., Noguchi, E., Yoshihara, S., Ebisawa, M., Saito, H., Matsumoto, K., Nakamura, Y., Ziegler. S.F. and Tamari M. Functional analysis of the Thymic Stromal Lymphopoietin Variants in Human Bronchial Epithelial Cells. Am. J. Respir. Cell

Mol. Biol. 40: 368-374, 2009.
50. Sakashita, M., Yoshimoto, T., Hirota, T., Harada, M., Okubo, K., Osawa, Y., Fujieda, S., Nakamura, Y., Yasuda, K., Nakanishi, K. and Tamari, M. Association of serum IL-33 level and the IL-33 genetic variant with Japanese cedar pollinosis. Clin. Exp. Allergy. 38: 1875-1881, 2008.
51. Hirata, D., Yamabuki, T., Miki, D., Ito, T., Tsuchiya, E., Fujita, M., Hosokawa, M., Chayama, K., Nakamura, Y. and Daigo, Y. Involvement of epithelial cell transforming sequence-2 oncoantigen in lung and esophageal cancer progression. Clin. Cancer Res. 15: 256-266, 2009.
52. Dobashi, S., Katagiri, T., Hirota, E., Ashida, S., Daigo, Y., Shuin, T., Fujioka, T., Miki, T. and Nakamura Y. Involvement of TMEM22 overexpression in the growth of renal cell carcinoma cells. Oncol. Rep. 21: 305-312, 2009.
53. Zembutsu, H., Suzuki, Y., Sasaki, A., Tsunoda, T., Okazaki, M., Yoshimoto, M., Hasegawa, T., Hirata, K. and Nakamura Y. Predicting response to Docetaxel neoadjuvant chemotherapy for advanced breast cancers through genome-wide gene expression profiling. Int. J. Oncol. 34: 361-370, 2009.
54. Nakamura, Y. DNA variations in human and medical genetics: 25 years of my experience. (review) J. Hum. Genet. 54: 1-8, 2009.
55. Ozaki, K., Sato, H., Inoue, K., Tsunoda, T., Sakata, Y., Mizuno, H., Lin, T.-H., Miyamoto, Y., Aoki, A., Onouchi, Y., Sheu, S.-H., Ikegawa, S., Odashiro, K., Nobuyoshi, M., Juo, S.-H. H., Hori, M., Nakamura, Y. and Tanaka T.A functional variation in BRAP confers risk of myocardial infarction in Asian populations. Nat. Genet., in press, 2009.
56. Kashiwaya, K. Hosokawa, M., Eguchi, H., Ohigashi, H., Ishikawa, O., Shinomura, Y., Nakamura, Y. and Nakagawa H. Identification of C2orf18, Termed ANT2BP (ANT2-binding protein), as one of key molecules involved in pancreatic carcinogenesis. Cancer Sci. 100: 457-464, 2009.
57. Nagayama, S., Yamada, E., Kohno, Y., Aoyama, T., Fukukawa, C., Kubo, H., Watanabe, G., Katagiri, T., Nakamura, Y., Sakai, Y. and Toguchida J. Inverse correlation of the upregulation of FZD10 expression and the activation of β-catenin in synchronous colorectal tumors. Cancer Sci., in press, 2009.
58. Ueda, K., Fukase, Y., Katagiri, T., Ishikawa, N. Irie, S., Sato, T., Ito, H., Nakayama, H., Miyagi, Y., Tsuchiya, E., Kohno, N., Shiwa, M., Nakamura, Y. and Daigo, Y. Targeted glycoproteomics for the discovery of lung cancer-associated glycosylation disorders using lectin-coupled ProteinChip arrays. Proteomocs., in press, 2009.
59. The International Warfarin Pharmacogenetics Consortium: Improved warfarin dosing with a global pharmacogenetic algorithm. N. Engl. J. Med. 360: 753-764, 2009.
60. Betcheva, E.T., Mushiroda, T., Takahashi, A., Kubo, M., Karachanak, S.K., Zaharieva, I.T., Vazharov,a R.V., Dimova, I.I., Milanova, V. K., Tolev, T., Kirov, G., Owenm M,J,, O'Donovanm M,C,, Kamatanim N, Nakamura, Y. and Toncheva, D.I. Case-control association study of 59 candidate genes reveals the DRD2 SNP rs6277 (C957T) as the only susceptibility factor for schizophrenia in Bulgarian population. J. Hum. Genet. 54: 98-107, 2009.
61. Fukukawa, C., Nagayama, S., Tsunoda, T., Toguchida, J., Nakamura, Y. and Katagiri, T. Activation of non-canonical Dvl-Rac1-JNK pathway by Frizzled-homologue 10 (FZD10) in human synovial sarcoma. Oncogene., in press, 2009.
62. Yosifova, A., Mushiroda, T., Stoianov, D., Vazharova, R., Dimova, I., Karachanak, S., Zaharieva, I., Milanova, V., Madjirova, N., Gerdjikov, I., Tolev, T., Velkova, S., Kirov, G., Owen, M.J., O'Donovan, M.C., Toncheva, D. and Nakamura, Y. Case-control association study of 65 candidate genes revealed a possible association of a SNP of HTR5A to be a factor susceptible to bipolar disease in Bulgarian population. J. Affective Disorders., in press, 2009.
63. Kamatani, Y., Wattanapokayakit, S., Ochi, H., Kawaguchi, T., Takahashi, A., Hosono, N., Kubo, M., Tsunoda, T., Kamatani, N., Kumada, H., Puseenam, A., Sura, T., Daigo, Y., Chayama, K., Chantratita, W., Nakamura, Y. and Matsuda, K. Identification of association of genetic variations in HLA-DP locus with chronic hepatitis B in Asian population through genome-wide association study. Nat. Genet., in press, 2009.
64. Tamura, K., Furihata, M., Chung, S., Uemura, M., Yoshioka, H., Iiyama, T., Ashida, S., Nasu, Y., Fujioka, T., Shuin, T., Nakamura, Y. and Nakagawa, H. Stanniocalcin 2 (STC2) over-expression in castration-resistant prostate cancer and aggressive prostate cancer. Cancer Sci., in press, 2009.
65. Tsukada, H., Ochi, H., Maekawa, T., Abe, H., Fujimoto, Y., Tsuge, M., Takahashi, H., Kumada, H., Kamatani, N., Nakamura, Y. and Chayama, K.: Hiroshima Liver Study Group; Toranomon Hospital. A Polymorphism in MAPKAPK3 affects response to in-

terferon therapy for chronic hepatitis C. Gastroenterology., in press, 2009.
66. Dunleavy, E.M., Roche, D., Tagami, H., Lacoste, N., Ray-Gallet, D., Nakamura, Y., Daigo, Y., Nakatani, Y. and Almouzni-Pettinotti, G. HJURP, a key CENP-A-partner for maintenance and deposition of CENP-A at centromeres at late telophase/G1. Cell., in press, 2009.

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. Test statistics correction for data of structured population

Because the genetic epidemiology studies on complex genetic traits target relatively weak factors, 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 modify the genomic control method so that it could correct the Fisher's exact test statistics.

b. Characterization of exact 2×3 test for SNP case-control association test data

The 2×3 contingency table test of SNP data is the basic unit of genome-wide association studies. We investigate the factors to affect the dis-

crepancy between the asymptotic test and the exact test for 2×3 contingency tables.

c. Geometric evaluation of SNP contingency table tests.

The 2×3 SNP contingency table tests are described in the context of geometry and characterize various tests for 2×3 tables and define tests fit for biological models by interpreting tables in the context of geometry.

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.

b. Geometric expression of haplotype populations

Haplotypes are consisted of alleles of multiple markers. We attempt 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 Autoimmune Diseases, CGM, 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

Gotoh N, Yamada R, Matsuda F, Yoshimura N, and Iida T. Manganese Superoxide Dismutase Gene (SOD2) Polymorphism and Exudative Age-related Macular Degeneration in the Japanese Population. Am J Ophthalmol 146: 146, 2008

Nakayama-Hamada M, Suzuki A, Furukawa H, Yamada R, and Yamamoto K. Citrullinated fibrinogen inhibits thrombin-catalyzed fibrin polymerization. J Biochem 144: 393-8, 2008

Okada Y, Mori M, Yamada R, Suzuki A, Kobayashi K, Kubo M, Nakamura Y, and Yamamoto K. SLC22A4 Polymorphism and Rheumatoid Arthritis Susceptibility: A Replication Study in a Japanese Population and a Metaanalysis. J Rheumatol 35: 1273-8, 2008

Shimane K, Kochi Y, Yamada R, Okada Y, Suzuki A, Miyatake A, Kubo M, Nakamura Y, and Yamamoto K. A single nucleotide polymorphism in the IRF5 promoter region is associated with susceptibility to rheumatoid arthritis in the Japanese patients. Ann Rheum Dis. (in press)

Suzuki A, Yamada R, Kochi Y, Sawada T, Okada Y, Matsuda K, Kamatani Y, Mori M, Shimane K, Hirabayashi Y Takahashi A, Tsunoda T, Miyatake A, Kubo M, Kamatani N, Nakamura Y, and Yamamoto K. Functional SNPs in CD244 increase the risk of rheumatoid arthritis in a Japanese population. Nat Genet 40: 1224-9, 2008

Yamada R. Primer: SNP-associated studies and what they can teach us. Nat Clin Pract Rheumatol 4: 210-7, 2008

Yamada R, and Okada Y. An optimal dose-effect mode trend test for SNP genotype tables. Genet Epidemiol 33: 114-27, 2009

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. Tissue and developmental stage specificity of *trans*-splicing in *C. intestinalis*

Nicolas Sierro, Shuang Li, Yutaka Suzuki[1], Riu Yamashita, and Kenta Nakai: [1]Graduate School of Frontier Sciences, U. Tokyo

Ciona intestinalis is a useful model organism to analyze chordate development and genetics. However, unlike vertebrates, it shares a unique mechanism called *trans*-splicing with lower eukaryotes. Our computational analysis of *trans*-splicing in *C. intestinalis* showed that although the amount of non-*trans*-spliced and *trans*-spliced genes is usually equivalent, the expression ratio between the two groups varies significantly with tissues and developmental stages. Among the seven tissues studied, the observed ratios ranged from 2.53 in "gonad" to 19.53 in "endostyle", and during development they increased from 1.68 at the "egg" stage to 7.55 at the "juvenile" stage. We hypothesize that this enrichment in *trans*-spliced mRNAs in early developmental stages might be related to the abundance of *trans*-spliced mRNAs in "gonad". To further investigate this phenomenon, we are currently analyzing a larger set of short 5′-EST tags obtained from specific tissues and developmental stages.

2. Improvement of the database of tunicate gene regulation

Nicolas Sierro, Takehiro Kusakabe[2], Yutaka Suzuki[1], Riu Yamashita and Kenta Nakai: [2] University of Hyogo

The database of tunicate gene regulation, DBTGR, was first released in 2006 as a small database summarizing published information about tunicate promoters and *cis*-regulatory regions. In 2008 it was extended to include gene expression reporter constructs as well as a new genome browser providing all whole genome alignments between *Ciona intestinalis* and *Ciona savignyi*. The description of 81 gene expression reporter vectors, as well as sample images of the expression observed with them in *Ciona* is now available, and the database provides users with contact information to the owners of these constructs. With the new flexible genome browser built in DBTGR, users have now access to two different genome alignments between *C. intestinalis* and *C. savignyi*, obtained with different algorithms. In addition, predicted binding sites for the JASPAR core matrices, as well as regulatory

elements and binding sites reported in literature are also directly available. DBTGR is accessible at http://dbtgr.hgc.jp.

3. Promoter architecture analysis and prediction of expression

Alexis Vandenbon and Kenta Nakai

Regulation of transcription is implemented through transcription factors (TFs) binding regulatory regions in the neighborhood of genes. We can make the assumption that genes showing similar expression profiles contain some shared structural patterns in their regulatory regions. Until recently, these patterns were considered only on the level of presence or absence of specific transcription factor binding sites (TFBSs), but there is growing evidence that additional structural patterns exist. Here we are focusing our attention not only on the presence of TFBSs, but also on their orientation and positioning with regard to the transcription start site and also between pairs of TFBSs. We developed an approach for extracting such structural motifs from promoter sequences and subsequently combining them to make a promoter structure model. We applied our model on a dataset of promoter sequences of muscle-specific genes of *Caenorhabditis elegans* and verified that our model is capable of distinguishing muscle-expressed genes from genes not expressed in muscle tissues based on the structure of their regulatory regions. We are further developing our model, and runs on *Mus musculus* datasets indicate that the approach is applicable in mammals too.

4. Characterization and definition of promoter-associated CpG islands in ascidian genomes

Kohji Okamura, Riu Yamashita, Koki Nishitsuji[2], Yutaka Suzuki[1], Takehiro Kusakabe[2], and Kenta Nakai

While CpG islands are often linked to a promoter in mammals, their existence in invertebrates is unclear. Since there is a striking difference in DNA methylation pattern between vertebrates and invertebrates, which show global and fractional methylation, respectively, the function of methylation *per se* in the latter group is also elusive. To address these questions, we performed determination of TSSs of ascidian genes by combination of the oligo-capping method and massive-scale cDNA sequencing. As a result, we found characteristic features of ascidian promoters. They tend to be G+C- and CpG-rich, but over a narrower range around the TSSs. Furthermore, almost all promoters fall into the same category, whereas vertebrate promoters are divided into two classes in terms of CpG. Comparison of the experimental result with the genome of another ascidian species also supported our finding, leading to the first definition of promoter-associated CpG islands in invertebrate organisms.

5. Computational verifications of gene regulatory networks in ascidian early development

Xuyang Yuan, Atsushi Kubo[3], Yutaka Satou[3], and Kenta Nakai: [3]Kyoto University

The ascidian *Ciona intestinalis* has been useful as a model system to explore chordate development. Systematic gene knockdown experiments highly contributed to the depiction of the gene regulatory network governing ascidian early development. However, limitations of the experiment itself prevent the blueprint from giving further information regarding direct or indirect regulation. In this study, we are computationally detecting direct target genes of each transcription factor by scanning all promoter sequences for its binding site. For representing the sequence specificity of transcription factors, we utilized positional weight matrices, of which threshold values we need to set. We maximized an over-representation index (ORI) value to find the optimum threshold. For *trans*-acting factors whose binding sites are unknown but have orthologues with known binding sites, we are predicting them by the examination of orthologues. The regulation network of *C. intestinalis* transcription factor *ZicL* is consistent with the data of a newly produced ChIP-chip experiment. Using our method together with ChIP-chip data, we further expanded the original network to cover all 16,000 *C. intestinalis* genes. So that not only the kernel components of the regulatory network making body plan, but also peripheral components which actually make building block of the body are included.

6. Pseudocounts for transcription factor binding sites

Keishin Nishida, Martin Frith[4], and Kenta Nakai: [4]CBRC, AIST

To represent the sequence specificity of transcription factors, the position weight matrix (PWM) is widely used. In most cases, each element is defined as a log likelihood ratio of a base appearing at a certain position, which is es-

timated from a finite number of known binding sites. To avoid bias due to this small sample size, a certain numeric value, called a pseudocount, is usually allocated for each position, and its fraction according to the background base composition is added to each element. So far, there has been no consensus on the optimal pseudocount value. In this study, we simulated the sampling process by artificially generating binding sites based on observed nucleotide frequencies in a public PWM database, and then the generated matrix with an added pseudocount value was compared to the original frequency matrix using various measures. Although the results were somewhat different between measures, in many cases, we could find an optimal pseudocount value for each matrix. These optimal values are independent of the sample size and are clearly anti-correlated with the information content of the original matrices, meaning that larger pseudocount vales are preferable for less conserved binding sites. As a simple representative, we suggest the value of 0.8 for practical uses.

7. Definition and analysis of alternative promoters using a huge number of TSS information

Riu Yamashita, Yutaka Suzuki[1], Hiroyuki Wakaguri[1], Sumio Sugano[1], Kenta Nakai

In order to support transcriptional studies, we have constructed a database, DataBase of Transcriptional Start Sites (DBTSS: http://dbtss.hgc.jp), which includes a number of 5′-end sequences produced by oligo-capping method. Recently, we have added 296.5 million tags from eight kinds of cells (15 kinds of experimental conditions) using a SOLEXA sequencer. Here, we performed analysis of alternative promoters with these data. From these data, we obtained 75,918 promoters. These promoters could be classified into 36,251 gene regions and 39,667 intergenic regions. Former intragenic promoters corresponded to 14,307 genes ,and 5,428 of them have one promoter, and 8,879 genes have more than one promoter. For each gene, we defined the promoter with the largest number of tags as the '1st promoter' and the 2nd highest promoter as the '2nd promoter'. Between different cell types, the average percentage of the discrepancy for 1st and 2nd promoters was 28.3%. On the other hand, we observed 9.6% of difference for promoters expressed in the same cell types with different conditions. These results indicate that the expression ratio of promoters is conserved among cells. We also observed that 2nd promoters preferentially occur in downstream regions of 1st promoters.

8. Effects of Alu elements on global nucleosome positioning in the human genome

Yoshiaki Tanaka, Riu Yamashita and Kenta Nakai

Because chromatin can limit the accessibility of regulatory sites, understanding the genome sequence-specific positioning of nucleosome is important for the analyses of transcription and replication. It has been previously reported that the 10-bp dinucleotide periodicities are strongly associated with nucleosome positioning, but it is unknown whether these features can affect in vivo nucleosome locations through the wholte genomes of all eukaryote. Fourier analysis to the genome fragments indicates that these are not common in 16 eukaryotes, but the two primate-specific periodicities (84-bp and 167-bp) are observed. The 167 bp is similar with the sum of the lengths of a nucleosome unit and its linker region. After masking Alu elements, these periodicities were greatly diminished. Therefore, we next analyzed the distribution of nucleosomes in the vicinity of them. Using two independent large-scale sets of recently published nucleosome mapping data, we found that (1) there are one or two fixed slot(s) for nucleosome positioning within the Alu element and (2) the positioning of neighboring nucleosomes seems to be in phase, more or less, with the presence of Alu elements. Our study provides an important clue to understanding the whole chromatin composition of the primate genomes.

9. Estimation and Comparison of minimal cellular function sets for bacteria and eukaryotes

Yusuke Azuma and Kenta Nakai

A minimal cell, containing only necessary and sufficient components, has been estimated mostly by the reduction of the genome of a living cell. But the "minimal gene set" obtained by the former approach may be inaccurate due to the effect of evolution. Thus, we tried to detect the minimal cellular function, instead. As cellular functions, we used KEGG pathway maps. The minimal pathway maps were detected as a combination of the conserved pathway maps and the organism-specific pathway maps. The conserved pathway maps are those containing more orthologous genes in all pathway maps and are estimated by homology searches. They should be close to the minimal pathways but it is not sure whether they are organized to sus-

tain life from only external nutrients like living cells. Then, the organism-specific pathway maps are detected as those that can synthesize compounds required for the conserved pathway maps from nutrients. The minimal pathway maps detected for bacteria agree well with the experimental essential genes. Most of the catabolization pathways were selected as organism-specific pathways rather than conserved ones, suggesting that they are adapted to each environment. The minimal pathway maps of eukaryotes contain more pathway maps for DNA repair than those of bacteria. In addition, there are more links in the pathways of eukaryotes. Thus, it is likely that eukaryotes need to be more stable genetically.

10. Development of new indices to evaluate protein-protein interfaces: Assembling space volume, assembling space distance, and global shape descriptor

M. Maeda[5] and K. Kinoshita: [5]National Institute of Agrobiological Sciences

Protein-protein interaction is an initial step to realize complex biological functions, therefore, understanding of the protein-protein interfaces will give us a clue to predict the protein complex structures. For the purpose, efficient descriptors of the interface and database analyses are important. In this study, we developed three new descriptors of protein-protein interfaces, that is, assembling space volume, assembling space distance, and global shape descriptor, by using Delaunay tessellation technique. The first two indexes enable us to evaluate how well the protein interfaces are build up, and the third descriptor quantifies the complexity of the protein-protein interfaces. Systematic comparison with some existing descriptors, our indexes could elucidate the different aspects of the protein interfaces.

11. ATTED-II: a coexpression database for Arabidopsis.

T. Obayashi, S. Hayashi[6], M. Saeki[6], H. Ohta[6], K. Kinoshita: [6]Tokyo Institute of Technology

ATTED-II (http://atted.jp) is a database of gene coexpression in Arabidopsis that can be used to design a wide variety of experiments, including the prioritization of genes for functional identification or for studies of regulatory relationships. Here, we report updates of ATTED-II that focus especially on functionalities for constructing gene networks with regard to the following points: (i) introducing a new measure of gene coexpression to retrieve functionally related genes more accurately, (ii) implementing clickable maps for all gene networks for step-by-step navigation, (iii) applying Google Maps API to create a single map for a large network, (iv) including information about protein-protein interactions, (v) identifying conserved patterns of coexpression and (vi) showing and connecting KEGG pathway information to identify functional modules. With these enhanced functions for gene network representation, ATTED-II can help researchers to clarify the functional and regulatory networks of genes in Arabidopsis.

12. PiSite: a database of protein interaction sites using multiple binding states in the PDB

M. Higurashi, T. Ishida, and K. Kinoshita

The vast accumulation of protein structural data has now facilitated the observation of many different complexes in the PDB for the same protein. Therefore, a single protein complex is not sufficient to identify their interaction sites, especially for proteins with multiple binding states or different partners, such as hub proteins. Thus, we developed a database that provides protein-protein interaction sites at the residue level with consideration of multiple complexes at the same time, by mapping the binding sites of all complexes containing the same protein in the PDB. We also implemented easy web-interfaces with an interactive viewer working with typical web-browsers, and the different binding modes can be checked visually.

13. Discrimination between biological interfaces and crystal-packing contacts

Y. Tsuchiya, H. Nakamura[7] and K. Kinoshita: [7]Osaka University

The quaternary structures of proteins are the bases of their physiological functions, and thus it is indispensable to know the biologically relevant complexes of proteins to understand their functions at the molecular level. The structures of proteins are usually determined by X-ray crystallography, which could contain non-biological interactions due to the nature of crystals. Therefore, discrimination between biologically relevant interfaces and artificial crystal-packing contacts in crystal structures is required. We developed a discrimination method between biological and non-biological interfaces, which evaluates protein-protein interfaces in terms of complementarities for hydrophobicity,

electrostatic potential and shape on the protein surfaces, and chooses the most probable biological interfaces among all possible contacts in the crystal. Our discrimination method achieved a good success rate, comparable to that of the contact area-dependent discrimination. Subsequent detailed review of the discrimination results raised the success rate to 91.4%.

14. Effect of surface-to-volume ratio of proteins on hydrophilic residues

M. Shirota, T. Ishida and K. Kinoshita

The size of a protein has been shown to affect both the amino acid composition and the residue burial in the protein. To demonstrate that these effects are the results from the reduction of surface regions relative to the volume in larger proteins, we examined the effect of surface-to-volume ratio (SVR), which is the ratio between the accessible surface area and volume of a protein, to amino acid composition. The reduction of several hydrophilic residues was more strongly correlated with SVR than with protein size (*i.e*. the number of amino acids), which indicats that SVR directly affected the amino acid composition. Furthermore, these hydrophilic residues also increased in buried fraction at the same time of the reduction. The increase in burial was found to be accelerated compared with the decrease in occurrence as SVR decreased below SVR$=0.3\,\text{Å}^{-1}$ (approximately protein size exceeded 132 residues) except for lysine, which was the most difficult for being buried.

15. Prediction of disordered regions in proteins based on the meta approach

Takashi Ishida and Kengo Kinoshita

Intrinsically disordered regions in proteins have no unique stable structures without their partner molecules, thus these regions sometimes prevent high-quality structure determination. Furthermore, proteins with disordered regions are often involved in important biological processes, and the disordered regions are considered to play important roles in molecular interactions. Therefore, identifying disordered regions is important to obtain high-resolution structural information and to understand the functional aspects of these proteins. Thus, we developed a new prediction method for disordered regions in proteins based on the meta approach and implemented a web-server for this prediction method. The method predicts the disorder tendency of each residue using support vector machines from the prediction results of the seven independent predictors. As a result of our evaluation, the meta approach achieved higher prediction accuracy than previously developed methods.

16. A cavity with an appropriate size is the basis of the PPIase activity

Teikichi Ikura[8], Kengo Kinoshita, Nobutoshi Ito[8]: [8]Tokyo Medical and Dental University

Peptidyl-prolyl isomerases (PPIase) are important enzymes in biological systems, but the catalytic mechanisms are not well understood. To elucidate the essential amino acids for the enzymatic activities, we have carried out the similarity search of atomic configurations of the active site of PPIase against the known protein structures, and found alpha amylase and prolyl endopeptidase have the similar spatial arrangement of atoms with PPIase active sites. Furthermore, we proved experimentally that these proteins actually have the PPIase activities, which have not been considered at all. In addition, we created the similar hole in the barnase, which is a enzyme to catalyze the ribonuclease activity and does not have the PPIase activities, and found that the mutated barnase exhibit the PPIase activity. These results indicate that the PPIase activity can be realized by a hole with appropriate size on the surface of protein.

17. COXPRESdb: co-expressed gene database for mouse and human.

T. Obayashi, S. Hayashi[6], M. Shibaoka[6], M. Saeki[6], H. Ohta[6], K. Kinoshita

A database of coexpressed gene sets can provide valuable information for a wide variety of experimental designs, such as targeting of genes for functional identification, gene regulation and/or protein-protein interactions. Coexpressed gene databases derived from publicly available GeneChip data are widely used in Arabidopsis research, but platforms that examine coexpression for higher mammals are rather limited. Therefore, we have constructed a new database, COXPRESdb (coexpressed gene database) (http://coxpresdb.hgc.jp), for coexpressed gene lists and networks in human and mouse. Coexpression data could be calculated for 19 777 and 21 036 genes in human and mouse, respectively, by using the GeneChip data in NCBI GEO. COXPRESdb enables analysis of the four types of coexpression networks: (i) highly coexpressed genes for every gene, (ii) genes with the same GO annotation, (iii) genes expressed in the

same tissue and (iv) user-defined gene sets. 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. COXPRESdb also provides a view to compare the human and mouse coexpression patterns to estimate the conservation between the two species.

18. Influence of proteins and cholesterol on biological membranes analyzed by molecular dynamics

Naoya Fujita, Takashi Ishida and Kengo Kinoshita

Protein-membrane interactions are fundamental for both protein functions and membrane properties. By means of these interactions, suitable configurations of membrane molecules can generate heterogeneity, such as lipid rafts and transportsome regions, in the membrane. To reveal the bidirectional influences between proteins and surrounding lipids, we performed molecular dynamics simulations of biological membranes with and without proteins and cholesterol, and compared those trajectories. As a result, alamethicin, a small transmembrane peptide, was shown to reduce the whole membrane undulation, in addition to decreasing local membrane thickness according to the size of alamethicin's hydrophobic region. On the contrary, water accessibility of alamethicin and its hydrogen bonds with lipids were different depending on the cholesterol availability. Further investigations with aquaporin are also being performed.

Publications

Chiba, H., Yamashita, R., Kinoshita, K. and Nakai, K. Weak correlation between sequence conservation in promoter regions and in protein-coding regions of human-mouse orthologous gene pairs. *BMC Genomics*. 9: 152, 2008.

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*. 36: D793-D799, 2008.

Hatada, I., Morita, S., Kimura, M., Horii, T., Yamashita, R. and Nakai, K. Genome-wide demethylation during neural differentiation of P19 embryonal carcinoma cells. *J. Human Genet*. 53 (2): 185-191, 2008.

Hatanaka, Y., Nagasaki, M., Yamaguchi, R., Obayashi, T., Numata, K., Imoto, S., Shimamura, T., Kinoshita, K., Nakai, K. and Miyano, S. A novel strategy to search concerted transcription factor activities using gene expression profile and genomic data. *Genome Informatics*. 20: 212-221, 2008.

Higurashi, M., Ishida, T. and Kinoshita, K. PiSite: a database of protein interaction sites using multiple binding states in the PDB. *Nucleic Acids Res*. 37: D360-364, 2009.

Ikura, T., Kinoshita, K. and Ito, N. A cavity with an appropriate size is the basis of the PPIase activity. *Protein Eng Des Sel*. 21: 83-89, 2008.

Ishida, T. and Kinoshita, K. Prediction of disordered protein regions based on meta-approach. *Bioinformatics*. 24: 1344-1348, 2008.

Maeda, M. and Kinoshita, K. Development of new indices to evaluate protein-protein interfaces: Assembling space volume, assembling space distance, and global shape descriptor. *J. Mol. Graph. Mod*. 27: 706-711, 2009.

Miura, K., Toh, H., Hirakawa, H., Sugii, M., Murata, M., Nakai, K., Tashiro, K., Kuhara, S., Azuma, Y. and Shirai, M. Genome-wide analysis of *Chlamydophila pneumoniae* gene expression at the late stage of infection. *DNA Res*. 15 (2): 83-91, 2008.

Murakami, K., Imanishi, T., Gojobori, T. and Nakai, K. Two different classes of co-occurring motif pairs found by a novel visualization method in human promoter regions. *BMC Genomics*. 9 (1): 112, 2008.

Nishida, K., Frith, M. and Nakai, K. Pseudocounts for transcription factor binding sites. *Nucl. Acids Res*. 37: 939-944, 2009; published online on December 23, 2008.

Obayashi, T., Hayashi, S., Shibaoka, M., Saeki, M., Ohta, H. and Kinoshita, K. COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res*. 36: D77-82, 2008.

Obayashi, T., Hayashi, S., Saeki, M., Ohta, H., and Kinoshita, K. ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic Acids Res*. 37: D987-991, 2009.

Okamura, K. and Nakai, K. Retrotransposition as a source of new promoters. *Mol. Biol. Evol*. 25 (6): 1231-1238, 2008.

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*. 36: D93-D96, 2008.

Sierro, N., Li, S., Suzuki, Y., Yamashita, R. and Nakai, K. Spatial and temporal preferences for trans-splicing in *Ciona intestinalis* revealed by

EST-based gene expression analysis. *Gene*. 430: 44-49, 2009; available online on October 21, 2008.

Shirota, M., Ishida, T. and Kinoshita, K. Effects of surface-to-volume ratio of proteins on hydrophilic residues, decrease in occurrence and increase in buried fraction. *Protein Sci*. 17: 1596-1602, 2008.

Tsuchihara, K., Suzuki, Y., Wakaguri, H., Irie, T., Tanimoto, K., Hashimoto, S., Matsushima, K., Mizushima-Sugano, J., Yamashita, R., Nakai, K., Bentley, D., Esumi, H. and Sugano, S. Massive transcriptional start site analysis of human genes in hypoxia cells. *Nucl. Acids Res*. in press.

Tsuchiya, Y., Nakamura, H. and Kinoshita, K. Discrimination between biological interfaces and crystal-packing contacts. *Compt Biol Chem*. 1: 99-113, 2008.

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*. 15 (1): 3-11, 2008.

Vandenbon, A. and Nakai, K. Using simple rules on presence and positioning of motifs for promoter structure modeling and tissue specific expression prediction. *Genome Informatics*. Edited by Arthur, J. and Ng, S.-K. (Imperial College Press, London). vol. 21, pp. 188-199, 2008.

Wakaguri, H., Yamashita, R., Suzuki, Y., Sugano, S. and Nakai, K., DBTSS: DataBase of Transcription Start Sites, progress report 2008. *Nucl. Acids Res*. 36: D97-D101, 2008.

Yamashita, R., Suzuki, Y., Takeuchi, N., Wakaguri, H., Ueda, T., Sugano, S. and Nakai, K. Comprehensive detection of human terminal oligo-pyrimidine (TOP) gene and analysis of their characteristics. *Nucl. Acids Res*. 36 (11): 3707-3715, 2008.

Kinoshita, K., Kono, H., and Yura, K. Prediction of molecular interactions from 3D-structures: from small ligands to large protein complexes. Edited by Bujnicki, J. (Wiley and Sons, USA). in printing, 2009.

伊倉貞吉，木下賢吾，伊藤暢聡．ペプチジルプロリルイソメラーゼの構造・機能相関．蛋白質核酸酵素．54：167−172，2009.

木下賢吾．立体構造からのタンパク質機能予測：現状と展望．遺伝子医学MOOK. 14号．in press

中井謙太，ポール・ホートン．第3章　3．アミノ酸配列に基づくタンパク質の細胞内局在予測．実験医学増刊　vol. 26(7)：1106−1112, 2008.

中井謙太．タンパク質のシステム生物学．猪飼・伏見・卜部・上野川・中村・浜窪編．タンパク質の事典．朝倉書店．575−578，2008.

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
Project Assistant Professor	Hyongoo Hong, Ph.D.	特任助教	学術博士	洪賢秀
Project Assistant Professor	Ayako Kamisato	特任助教	法学修士	神里彩子

Department of Public Policy works 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 stem cell research regarding homecare services for ALS in East Asia. We also supported for "BioBank Japan" project from ethical, legal and social standpoints and ended the first questionnaire survey. We held Sci/Art Café twice at the Medical Science Museum as one of the outreach activities.

1. A comparative political study on stem cell research and genetic testing in East Asia

Supported by Japan Bioindustry Association, we conducted a comparative study on research policy on stem cells to examine broader social and cultural agendas on industrialization of stem cell research and genetic testing. We've interviewed main players in this area; the relevant authorities, bioindustry CEOs, 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 revised and the government seeks more strict regulation towards life science and healthcare. We've found some correlations in political options on stem cell research and genetic testing in terms of regulations among in East Asia.

2. Establishment of Office of Research Ethics (ORE)

Under the Dean's courageous decision, the IMSUT have established the Office of Research Ethics (ORE) for supporting research activities. Our department has main responsibility for managing the ORE and our research ethics review system, supported by Professor Hiroshi Kiyono of Division of Mucosal Immunology, Professor Kensuke Miyake of Division of Infectious Genetics, Professor Fumitaka Nagamura and Dr. Makiko Tajima of Department of Clinical Trial Safety Management, Professor Yasushi Kodama of Graduate School of Public Policy and Professor Akira Akabayashi of Graduate School of Medicine. After conducting our survey on past ethical reviews and a comparative study on research ethics review system in the US, the UK and South Korea, we checked our current problems which tend to stuck fluent research review process so as to secure quality assurance of ethical discussions. Since February 3rd of 2009, Ayako Kamisato has assumed main responsibility on "bench consulting" regarding consent, research protocols and pre-review on research ethics of all research involving human subjects. We will start communication with other relevant divisions on research ethics review founded by re-

search institutes and prepare for new study on research ethics review and ethical governance for future.

3. 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 free and fully informed consent from participants. We conducted our questionnaire survey to participants of the BioBank Japan Project. Our data shows that the younger participants thought that their personal analyzed data should be disclosed. The consent process had been well-worked out in advance and is fully 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 for 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. We've found that participants who responded that they had forgotten the whole consent process are not the elderly population. Furthermore, 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" three times in 2007 for MCs and participants. 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.

4. Sci/Art Café

According to the 3rd Science and Technology Basic Plan (FY2006-FY2010), outreach activities are promoted that aim for the sharing of public needs through interactive communication between researchers and the public. As one of such outreach activities, we held our original science café series called as "Sci/Art Café" twice in 2008. Our original intent of "Sci/Art Café" is to promote communication between scientists and those who don't have regular communication with science but love art. The 1st session called "Rhythm generated by network" was held in Shibuya during the 3rd World Rhythm Summit, supported by Dr. Atsuko Takamatsu (Waseda Univ.), Dr. Shin-ichi Nakagawa (RIKEN) and Dr. Hideaki Takeuchi (UT). The 2nd session called "Doing science, doing art" was held on October 8th at the Medical Science Museum in the IMSUT, supported by Dr. Hideo Iwasaki (Waseda Univ.) and Dr. Yoichiro Murakami (JST). We prepare for the 3rd session in next early summer 2009.

Publications

1. Ishiyama I, Nagai A, Muto K, Tamakoshi A, Kokado M, Mimura K, Tanzawa T, Yamagata Z. Relationship between Public Attitudes toward Genomic Studies Related to

Medicine and Their Level of Genomic Literacy in Japan. American Journal of Medical Genetics. 146A (13). 696-706, 2008.
2. 洪賢秀．韓国社会における子どもの「性保護」と性犯罪防止対策．比較法研究．70号．2009．印刷中．
3. 神里彩子・成澤光編著．生殖補助医療　生命倫理と法―基本資料集３．信山社．21–123，262–308，2008．
4. 張瓊方．諸外国における生殖補助医療の規制状況と実施状況(台湾)．生殖補助医療　生命倫理と法―基本資料集３．神里彩子・成澤光編．信山社．323–334,2008.
5. 大上泰弘・神里彩子・城山英明．イギリス及びアメリカにおける動物実験規制の比較分析―日本の規制体制への示唆．社会技術研究論文集５号．132–142，2008．
6. 大上泰弘・成廣孝・神里彩子・城山英明・打越綾子．日本における生命科学・技術者の動物実験に関する意識―生命科学実験及び動物慰霊祭に関するアンケート調査の分析．ヒトと動物の関係学会誌20号．66–73，2008．
7. 大上泰弘・神里彩子・城山英明．イギリスにおける動物の実験規制を支えている思考様式．科学技術社会論研究５号．84–92，2008．
8. 渡部麻衣子，上田昌文．人の必要を充足する科学技術．福祉工学における開発現場の分析．科学技術社会研究．138–151，2008．
9. 武藤香織．「脱医療化」する予測的な遺伝学的検査への日米の対応―遺伝病から栄養遺伝学的検査まで―．日米の医療―制度と倫理．杉田米行編．大阪大学出版会．203–224，2008．
10. 武藤香織．DNA親子鑑定は「ふしだらな」女性にとっての救済策か？．ジェンダー研究のフロンティア第４巻　テクノ／バイオ・ポリティクス―科学・医療・技術のいま．舘かおる編．作品社．238–264，2008．
11. 洪賢秀．研究用卵子提供の何が問題なのか―韓国黄禹錫論文捏造事件を中心に―．ジェンダー研究のフロンティア第４巻　テクノ／バイオ・ポリティクス―科学・医療・技術のいま．舘かおる編．作品社．196–214，2008
12. 張瓊方．生殖技術と台湾社会．ジェンダー研究のフロンティア第４巻　テクノ／バイオ・ポリティクス―科学・医療・技術のいま．舘かおる編．作品社．215–222，2008
13. 三村恭子，小門穂，武藤香織，張瓊方，洪賢秀，柘植あづみ．女性にやさしい機械のつくられ方―内診台を例にして．ジェンダー研究のフロンティア第４巻　テクノ／バイオ・ポリティクス―科学・医療・技術のいま．舘かおる編．作品社．223–240，2008
14. 神里彩子．生殖補助医療をめぐる議論―その回顧と展望―．家永登編『生殖技術と家族Ⅱ』．早稲田大学出版部．42–71，2008．
15. 渡部麻衣子，上田昌文編訳．エンハンスメント論争．身体・精神の増強と先端科学技術．社会評論社，2008．