Human Genome Center

Laboratory of DNA Information Analysis
Laboratory of Sequence Analysis
Laboratory of Genome Database

DNA情報解析分野
シークエンスデータ情報処理分野
ゲノムデータベース分野

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
Associate Professor	Rui Yamaguchi, Ph.D.	准教授	博士(理学)	山　口　　類
Assistant Professor	Yuichi Shiraishi, Ph.D.	助　教	博士(統計科学)	白　石　友　一
Project Assistant Professor	Atsushi Niida, Ph.D.	特任助教	博士(理学)	新井田　厚　司
Associate Professor	Tetsuo Shibuya, Ph.D.	准教授	博士(理学)	渋　谷　哲　朗

We are facing with biomedical big data comprising of ultra-high dimensional ultra-heterogeneous data. Our current mission is to develop computational/informatics strategy for medical informatics to implement personalized genomic medicine through genomics, systems biology and supercomputer.

1. Systems Cancer Research and Systems Biology

a. Recursive Random Lasso (RRLasso) for Identifying Anti-Cancer Drug Targets

Park H, Imoto S, Miyano S

Uncovering driver genes is crucial for understanding heterogeneity in cancer. L1-type regularization approaches have been widely used for uncovering cancer driver genes based on genome-scale data. Although the existing methods have been widely applied in the field of bioinformatics, they possess several drawbacks: subset size limitations, erroneous estimation results, multicollinearity, and heavy time consumption. We introduce a novel statistical strategy, called a Recursive Random Lasso (RRLasso), for high dimensional genomic data analysis and investigation of driver genes. For time-effective analysis, we consider a recursive bootstrap procedure in line with the random lasso. Furthermore, we introduce a parametric statistical test for driver gene selection based on bootstrap regression modeling results. The proposed RRLasso is not only rapid but performs well for high dimensional genomic data analysis. Monte Carlo simulations and analysis of the "Sanger Genomics of Drug Sensitivity in Cancer dataset from the Cancer Genome Project" show that the

proposed RRLasso is an effective tool for high dimensional genomic data analysis. The proposed methods provide reliable and biologically relevant results for cancer driver gene selection.

b. High performance computing of a fusion gene detection pipeline on the K computer

Ito S, Shiraishi Y, Shimamura T, Chiba K, Miyano S.

Recently developed high-throughput sequencers can generate a huge amount of omics data, and TOP500-class supercomputers are required to analyze such large datasets. However, these supercomputers do not usually support grid engines, which are commonly used in bioinformatics, making it necessary to parallelize the software. Parallelization and optimization require domain and specific knowledge, which poses a challenge for most bioinformaticians. We here propose a simple methodology for the parallelization of pipeline software. To demonstrate the efficacy of the methodology, we employed the Genomon-fusion as a sample software pipeline and ported it onto the K computer by applying our method. Simultaneous analysis of a massive amount of samples was performed using the K computer in a very short period of time.

c. Integrative Clustering of Cancer Genome Data using Infinite Relational Models

Chikahara Y, Niida A, Yamaguchi R, Imoto S, Miyano S

From the late 90's until today, the advances in highthroughput measurement technologies are remarkable and producing a huge amount of cancer genomic data. Due to the complexity of data, however, we have not still got a fully integrated view of genetic and transcriptional changes that differ among individuals. To visualize the differences in genetic and transcriptional data among patient samples, we focus on grouping of three types of features, i.e., genes, patient samples, and expression modules. We propose an integrative framework based on the biclustering of multiple types of biological data, i.e., copy number, gene expression, and module activity, by extending the Infinite Relational Models (IRM), a non-parametric Bayesian model used to perform a biclustering of binary data, for continuous data. We demonstrate a utility of the model using a colorectal cancer (CRC) dataset. Our result discovers a clinical insight that the activity of modules related to an immune system is associated with CRC patients survival, which demonstrates the ability of our novel integrative approach to group not only genes and modules but also patient samples based on their genetic and transcriptional alterations.

d. Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to time-dependent biomarker detection

Kayano M[1], Matsui H[2], Yamaguchi R, Imoto S, Miyano S: [1]Obihiro University of Agriculture and Veterinary Medicine, [2]Faculty of Mathematics, Kyushu University

High-throughput time course expression profiles have been available in the last decade due to developments in measurement techniques and devices. Functional data analysis, which treats smoothed curves instead of originally observed discrete data, is effective for the time course expression profiles in terms of dimension reduction, robustness, and applicability to data measured at small and irregularly spaced time points. However, the statistical method of differential analysis for time course expression profiles has not been well established. We propose a functional logistic model based on elastic net regularization (F-Logistic) in order to identify the genes with dynamic alterations in case/control study. We employ a mixed model as a smoothing method to obtain functional data; then F-Logistic is applied to time course profiles measured at small and irregularly spaced time points. We evaluate the performance of F-Logistic in comparison with another functional data approach, i.e. functional ANOVA test (F-ANOVA), by applying the methods to real and synthetic time course data sets. The real data sets consist of the time course gene expression profiles for long-term effects of recombinant interferon [Formula: see text] on disease progression in multiple sclerosis. F-Logistic distinguishes dynamic alterations, which cannot be found by competitive approaches such as F-ANOVA, in case/control study based on time course expression profiles. F-Logistic is effective for time-dependent biomarker detection, diagnosis, and therapy.

e. Elevated β-catenin pathway as a novel target for patients with resistance to EGF receptor targeting drugs

Nakata A[3], Yoshida R[4], Yamaguchi R, Yamauchi M[5], Tamada Y, Fujita A, Shimamura T, Imoto S, Higuchi T[4], Nomura M[6], Kimura T[7], Nokihara H[8], Higashiyama M[9], Kondoh K[10], Nishihara H[11], Tojo A[5], Yano S[3], Miyano S, Gotoh N[3]: [3]Kanazawa University, [4]The Institute of Statistical Mathematics, [5]Division of Molecular Therapy, Institute of Medical Science, University of Tokyo, [6]Tokyo Medical University, [7]Osaka City University, [8]National Cancer Center Hospital, [9]Osaka Medical Center for Cancer and Cardiovascular Diseases, [10]The Univer-

sity of Tokushima, [11]Hokkaido University

There is a high death rate of lung cancer patients. Epidermal growth factor receptor tyrosine kinase inhibitors (EGFR-TKIs) are effective in some lung adenocarcinoma patients with EGFR mutations. However, a significant number of patients show primary and acquire resistance to EGFR-TKIs. Although the Akt kinase is commonly activated due to various resistance mechanisms, the key targets of Akt remain unclear. Here, we show that the Akt-β-catenin pathway may be a common resistance mechanism. We analyzed gene expression profiles of gefitinib-resistant PC9M2 cells that were derived from gefitinib-sensitive lung cancer PC9 cells and do not have known resistance mechanisms including EGFR mutation T790M. We found increased expression of Axin, a β-catenin target gene, increased phosphorylation of Akt and GSK3, accumulation of β-catenin in the cytoplasm/nucleus in PC9M2 cells. Both knockdown of β-catenin and treatment with a β-catenin inhibitor at least partially restored gefitinib sensitivity to PC9M2 cells. Lung adenocarcinoma tissues derived from gefitinib-resistant patients displayed a tendency to accumulate β-catenin in the cytoplasm. We provide a rationale for combination therapy that includes targeting of the Akt-β-catenin pathway to improve the efficacy of EGFR-TKIs.

f. An Integrative Analysis to Identify Driver Genes in Esophageal Squamous Cell Carcinoma

Sawada G[12], Niida A, Hirata H[13], Komatsu H[12], Uchi R[13], Shimamura T, Takahashi Y[12], Kurashige J[13], Matsumura T[12], Ueo H, Takano Y, Ueda M[12], Sakimura S[13], Shinden Y[13], Eguchi H[13], Sudo T[13], Sugimachi K[13], Yamasaki M[12], Tanaka F[13], Tachimori Y[8], Kajiyama Y[14], Natsugoe S[15], Fujita H[16], Tanaka Y[17], Calin G, Miyano S, Doki Y[12], Mori M[12], Mimori K[13]: [12]Osaka University, [13]Kyushu University Beppu Hospital, [14]Juntendo University School of Medicine, [15]Kagoshima University School of Medicine, [16]Kurume University School of Medicine, [17]Saitama Cancer Center, [18]The University of Texas MD Anderson

Few driver genes have been well established in esophageal squamous cell carcinoma (ESCC). Identification of the genomic aberrations that contribute to changes in gene expression profiles can be used to predict driver genes. We searched for driver genes in ESCC by integrative analysis of gene expression microarray profiles and copy number data. To narrow down candidate genes, we performed survival analysis on expression data and tested the genetic vulnerability of each genes using public RNAi screening data. We confirmed the results by performing RNAi experiments and evaluating the clinical relevance of candidate genes in an independent ESCC cohort. We found 10 significantly recurrent copy number alterations accompanying gene expression changes, including loci 11q13.2, 7p 11.2, 3q26.33, and 17q12, which harbored CCND1, EGFR, SOX2, and ERBB2, respectively. Analysis of survival data and RNAi screening data suggested that GRB7, located on 17q12, was a driver gene in ESCC. In ESCC cell lines harboring 17q12 amplification, knockdown of GRB7 reduced the proliferation, migration, and invasion capacities of cells. Moreover, siRNA targeting GRB7 had a synergistic inhibitory effect when combined with trastuzumab, an anti-ERBB2 antibody. Survival analysis of the independent cohort also showed that high GRB7 expression was associated with poor prognosis in ESCC. Our integrative analysis provided important insights into ESCC pathogenesis. We identified GRB7 as a novel ESCC driver gene and potential new therapeutic target.

g. A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures

Shiraishi Y, Tremmel G, Miyano S, Stephens M[19]: [19]University of Chicago

Recent advances in sequencing technologies have enabled the production of massive amounts of data on somatic mutations from cancer genomes. These data have led to the detection of characteristic patterns of somatic mutations or "mutation signatures" at an unprecedented resolution, with the potential for new insights into the causes and mechanisms of tumorigenesis. Here we present new methods for modelling, identifying and visualizing such mutation signatures. Our methods greatly simplify mutation signature models compared with existing approaches, reducing the number of parameters by orders of magnitude even while increasing the contextual factors (e.g. the number of flanking bases) that are accounted for. This improves both sensitivity and robustness of inferred signatures. We also provide a new intuitive way to visualize the signatures, analogous to the use of sequence logos to visualize transcription factor binding sites. We illustrate our new method on somatic mutation data from urothelial carcinoma of the upper urinary tract, and a larger dataset from 30 diverse cancer types. The results illustrate several important features of our methods, including the ability of our new visualization tool to clearly highlight the key features of each signature, the improved robustness of signature inferences from small sample sizes, and more detailed inference of signature characteristics such as strand biases and sequence context effects at the base two positions 5′ to the mutated site. The overall framework of our work is based

on probabilistic models that are closely connected with "mixed-membership models" which are widely used in population genetic admixture analysis, and in machine learning for document clustering. We argue that recognizing these relationships should help improve understanding of mutation signature extraction problems, and suggests ways to further improve the statistical methods. Our methods are implemented in an R package pmsignature (https://github.com/friend1ws/pmsignature) and a web application available at https://friend1ws.shinyapps.io/pmsignature_shiny/.

h. The AURKA/TPX2 axis drives colon tumorigenesis cooperatively with MYC

Takahashi Y[12], Sheridan P, Niida A, Sawada G[12], Uchi R[13], Mizuno H[20], Kurashige J[13], Sugimachi K[13], Sasaki S[21], Shimada Y[8], Hase K[22], Kusunoki M[23], Kudo S[24], Watanabe M[25], Yamada K[26], Sugihara K[27], Yamamoto H[12], Suzuki A[28], Doki Y[12], Miyano S, Mori M[12], Mimori K[13]: [20]Chugai Pharmaceutical Co., Ltd, [21]Omori Red Cross Hospital, [22]National Defense Medical College, [23]Mie University, [24]Showa University Northern Yokohama Hospital, [25]Kitasato University, [26]Takano Hospital, [27]Tokyo Medical and Dental University, [28]Medical Institute of Bioregulation, Kyushu University

The MYC oncogene has long been established as a central driver in many types of human cancers including colorectal cancer. However, the realization of MYC-targeting therapies remains elusive; as a result, synthetic lethal therapeutic approaches are alternatively being explored. A synthetic lethal therapeutic approach aims to kill MYC-driven tumors by targeting a certain co-regulator on the MYC pathway. We analyzed copy number and expression profiles from 130 colorectal cancer tumors together with publicly available datasets to identify co-regulators on the MYC pathway. Candidates were functionally tested by in vitro assays using colorectal cancer and normal fibroblast cell lines. Additionally, survival analyses were carried out on another 159 colorectal cancer patients and public datasets. Our in silico screening identified two MYC co-regulator candidates, AURKA and TPX2, which are interacting mitotic regulators located on chromosome 20q. We found the two candidates showed frequent co-amplification with the MYC locus while expression levels of MYC and the two genes were positively correlated with those of MYC downstream target genes across multiple cancer types. In vitro, the aberrant expression of MYC, AURKA and TPX2 resulted in more aggressive anchorage-independent growth in normal fibroblast cells. Furthermore, knockdown of AURKA or TPX2, or treatment with an AURKA-specific inhibitor effectively suppressed the proliferation of MYC-expressing colorectal cancer cells. Additionally, combined high expression of MYC, AURKA and TPX2 proved to be a poor prognostic indicator of colorectal cancer patient survival. Through bioinformatic analyses and experiments, we proposed TPX2 and AURKA as novel co-regulators on the MYC pathway. Inhibiting the AURKA/TPX2 axis would be a novel synthetic lethal therapeutic approach for MYC-driven cancers.

i. Binary Contingency Table Method for Analyzing Gene Mutation in Cancer Genome

Ayada E, Niida A, Hasegawa T, Miyano S, Imoto S.

Gene mutations are responsible for a large proportion of genetic diseases such as cancer. Hence, a number of computational methods have been developed to find loci subject to frequent mutations in cancer cells. Since normal cells turn into cancer cells through the accumulation of gene mutations, the elucidation of interactive relationships among loci has great potential to reveal the cause of cancer progression; however, only a few methods have been proposed for measuring statistical significance of pairs of loci that are co-mutated or exclusively mutated. In this study, we proposed a novel statistical method to find such significantly interactive pairs of loci by employing the framework of binary contingency tables. Using Markov chain Monte Carlo procedure, the statistical significance is evaluated by sampling null matrices whose marginal sums are equal to those of the input matrix. We applied the proposed method to mutation data of colon cancer patients and successfully obtained significant pairs of loci.

2. Oncoimmunology

a. A TCR sequence data analysis pipeline: Tcrip

Yamaguchi R, Imoto S, Miyano S

We developed a TCR sequence data analysis pipeline Tcrip that was developed for analyzing millions of cDNA-sequence reads generated by high-throughput sequencers. Using Tcrip, we can dissect TCR-α and TCR-β sequences into specific segments and estimate amounts and types of unique T cell clones to characterize T cell repertoires. It also allows us to analyze unmapped parts of reads in detail. Then it provides reasons for unmappability and clues to search novel exon candidate sequences.

b. Quantitative T cell repertoire analysis by deep cDNA sequencing of T cell receptor α and β chains using next-generation sequencing (NGS)

Fang H[29], Yamaguchi R, Liu X[29], Daigo Y[30], Yew PY[29], Tanikawa C[31], Matsuda K[31], Imoto S, Miyano S, Nakamura Y[29]: [29]Section of Hematology/Oncology, Department of Medicine; The University of Chicago, [30]Shiga University of Medical Science Hospital, [31]Laboratory of Genome Sequence Technology, Human Genome Center, The Institute of Medical Science, The University of Tokyo

Immune responses play a critical role in various disease conditions including cancer and autoimmune diseases. However, to date, there has not been a rapid, sensitive, comprehensive, and quantitative analysis method to examine T-cell or B-cell immune responses. Here, we report a new approach to characterize T cell receptor (TCR) repertoire by sequencing millions of cDNA of TCR α and β chains in combination with a newly-developed algorithm. Using samples from lung cancer patients treated with cancer peptide vaccines as a model, we demonstrate that detailed information of the V-(D)-J combination along with complementary determining region 3 (CDR3) sequences can be determined. We identified extensive abnormal splicing of TCR transcripts in lung cancer samples, indicating the dysfunctional splicing machinery in T lymphocytes by prior chemotherapy. In addition, we found three potentially novel TCR exons that have not been described previously in the reference genome. This newly developed TCR NGS platform can be applied to better understand immune responses in many disease areas including immune disorders, allergies, and organ transplantations.

c. Quantitative characterization of T-cell repertoire in allogeneic hematopoietic stem cell transplant recipients

Yew PY[29], Alachkar H[29], Yamaguchi R, Kiyotani K[29], Fang H[29], Yap KL[29], Liu HT[29], Wickrema A[29], Artz A[29], van Besien K[29], Imoto S, Miyano S, Bishop MR[29], Stock W[29], Nakamura Y[29]

Allogeneic hematopoietic stem cell transplantation (HSCT) is one of curative treatment options for patients with hematologic malignancies. Although GVHD mediated by the donor's T lymphocytes remains the most challenging toxicity of allo-HSCT, graft-versus-leukemia (GVL) effect targeting leukemic cells, has an important role in affecting the overall outcome of patients with AML. Here we comprehensively characterized the TCR repertoire in patients who underwent matched donor or haplo-cord HSCT using next-generation sequencing approach. Our study defines the functional kinetics of each TCRA and TCRB clone, and changes in T-cell diversity (with identification of CDR3 sequences) and the extent of clonal expansion of certain T-cells. Using this approach, our study demonstrates that higher percentage of cord-blood cells at 30 days after transplant was correlated with higher diversity of TCR repertoire, implicating the role of cord-chimerism in enhancing immune recovery. Importantly, we found that GVHD and relapse, exclusive of each other, were correlated with lower TCR repertoire diversity and expansion of certain T-cell clones. Our results highlight novel insights into the balance between GVHD and GVL effect, suggesting that higher diversity early after transplant possibly implies lower risks of both GVHD and relapse following the HSCT transplantation.

3. Applications of Genomon to Cancer Genome Analyses

All laboratory members and many collaborators

We have been developing various pipe lines for analyzing genome sequence data including RNA sequences. By collaborations with many cancer researchers, we contributed to sequence data analyses using the supercomputer at Human Genome Center and K computer at AICS, RIKEN. Due to the limit of space, we list up our contributed papers: 7-8, 11-14, 18-22, 25, 27, 29-30, 32, 35, 37, 40-41, 44-45, 48-50, 52-53, 56.

4. Algorithms for Bioinformatics

a. Malphite: A convolutional neural network and ensemble learning based protein secondary structure predictor

Li Y, Shibuya T

We developed a convolution neural networks (CNN) and ensemble learning based method, called Malphite, to predict protein secondary structures. Maphite has three sub-models: the 1st CNN, PSI-PRED and the 2nd CNN. The 1st CNN and PSI-PRED are used to predict the initial secondary structure based on the position specific scoring matrix generated from PSIBLAST. The 2nd CNN performs ensemble learning by combining the prediction result of the 1st CNN and PSI-PRED and generate the final predictions. Malphite achieved a Q3 score of 82.3% and 82.6% for independently built dataset of 400 and 538 proteins respectively, and 82.6% ten-fold-cross validated accuracy for a dataset of 3000 proteins. In addition, Malphite accomplished a remarkable Q3 score of 83.6% for 122 targets from CASP10 (Critical Assessment of protein Structure Prediction), surpassing any secondary structure prediction technique to date. For all four datasets, Malphite consistently makes 2% more accurate prediction than PSI-PRED, which is a significantly step towards the estimated upper limit of protein secondary structure prediction accuracy of

90%.e

Publications

1. Ayada E, Niida A, Hasegawa T, Miyano S, Imoto S. Binary Contingency Table Method for Analyzing Gene Mutation in Cancer Genome. *Lecture Notes in Computer Science*. 9096: 12-23, 2015.
2. Chapman CG, Yamaguchi R, Tamura K, Weidner J, Imoto S, Kwon J, Marino SR, Miyano S, Nakamura Y, Kiyotani K. Characterization of T Cell Receptor Repertoire in Inflamed Tissues of Crohn's Disease through Deep Sequencing. *Inflammatory Bowel Diseases*. In press
3. Chiba K, Shiraishi Y, Nagata Y, Yoshida K, Imoto S, Ogawa S, Miyano S. Genomon ITDetector: a tool for somatic internal tandem duplication detection from cancer genome sequencing data. *Bioinformatics*. 31(1): 116-118, 2015.
4. Chikahara Y, Niida A, Yamaguchi R, Imoto S, Miyano S. Integrative Clustering of Cancer Genome Data using Infinite Relational Models. *BICOB 2015*. 11-18, 2015.
5. Eita MA, Shibuya T, Shoukry AA. Locating controlling regions of neural networks using constrained evolutionary computation. *CEC 2015*. 1581-1588, 2015.
6. Fang H, Yamaguchi R, Liu X, Daigo Y, Yew PY, Tanikawa C, Matsuda K, Imoto S, Miyano S, Nakamura Y. Quantitative T cell repertoire analysis by deep cDNA sequencing of T cell receptor α and β chains using next-generation sequencing (NGS). *Oncoimmunology*. 3(12): e 968467, 2015.
7. Fujimoto A, Furuta M, Shiraishi Y, Gotoh K, Kawakami Y, Arihiro K, Nakamura T, Ueno M, Ariizumi S, Nguyen HH, Shigemizu D, Abe T, Boroevich KA, Nakano K, Sasaki A, Kitada R, Maejima K, Yamamoto Y, Tanaka H, Shibuya T, Shibata T, Ojima H, Shimada K, Hayami S, Shigekawa Y, Aikata H, Ohdan H, Marubashi S, Yamada T, Kubo M, Hirano S, Ishikawa O, Yamamoto M, Yamaue H, Chayama K, Miyano S, Tsunoda T, Nakagawa H. Whole-genome mutational landscape of liver cancers displaying biliary phenotype reveals hepatitis impact and molecular diversity. *Nat Commun*. 6: 6120, 2015.
8. Garg M, Nagata Y, Kanojia D, Mayakonda A, Yoshida K, Haridas Keloth S, Zang ZJ, Okuno Y, Shiraishi Y, Chiba K, Tanaka H, Miyano S, Ding LW, Alpermann T, Sun QY, Lin DC, Chien W, Madan V, Liu LZ, Tan KT, Sampath A, Venkatesan S, Inokuchi K, Wakita S, Yamaguchi H, Chng WJ, Kham SK, Yeoh AE, Sanada M, Schiller J, Kreuzer KA, Kornblau SM, Kantarjian HM, Haferlach T, Lill M, Kuo MC, Shih LY, Blau IW, Blau O, Yang H, Ogawa S, Koeffler HP. Profiling of somatic mutations in acute myeloid leukemia with FLT3-ITD at diagnosis and relapse. *Blood*. 126(22): 2491-2501, 2015.
9. Hasegawa T, Mori T, Yamaguchi R, Shimamura T, Miyano S, Imoto S, Akutsu T. Genomic data assimilation using a higher moment filtering technique for restoration of gene regulatory networks. *BMC Syst Biol*. 9: 14, 2015.
10. Hasegawa T, Niida A, Mori T, Shimamura T, Yamaguchi R, Miyano S, Akutsu T, Imoto S. A Likelihood-free Filtering Method via Approximate Bayesian Computation in Evaluating Biological Simulation Models. *Computational Statistics & Data Analysis*. 94: 63-74, 2016.
11. Hira A, Yoshida K, Sato K, Okuno Y, Shiraishi Y, Chiba K, Tanaka H, Miyano S, Shimamoto A, Tahara H, Ito E, Kojima S, Kurumizaka H, Ogawa S, Takata M, Yabe H, Yabe M. Mutations in the gene encoding the E2 conjugating enzyme UBE2T cause Fanconi anemia. *Am J Hum Genet*. 96(6): 1001-1007, 2015.
12. Hoshino A, Nomura K, Hamashima T, Isobe T, Seki M, Hiwatari M, Yoshida K, Shiraishi Y, Chiba K, Tanaka H, Miyano S, Ogawa S, Takita J, Kanegane H. Aggressive transformation of anaplastic large cell lymphoma with increased number of ALK-translocated chromosomes. *Int J Hematol*. 101(2): 198-202, 2015.
13. Huang D, Nagata Y, Grossmann V, Radivoyevitch T, Okuno Y, Nagae G, Hosono N, Schnittger S, Sanada M, Przychodzen B, Kon A, Polprasert C, Shen W, Clemente MJ, Phillips JG, Alpermann T, Yoshida K, Nadarajah N, Sekeres MA, Oakley K, Nguyen N, Shiraishi Y, Shiozawa Y, Chiba K, Tanaka H, Koeffler HP, Klein HU, Dugas M, Aburatani H, Miyano S, Haferlach C, Kern W, Haferlach T, Du Y, Ogawa S, Makishima H. BRCC3 mutations in myeloid neoplasms. *Haematologica*. 100(8): 1051-1057, 2015.
14. Ikenoue T, Yamaguchi K, Koumura M, Imoto S, Yamaguchi R, Shimizu E, Kasuya S, Shibuya T, Hatakeyama S, Miyano S, Furukawa Y. Attenuated familial adenomatous polyposis with desmoids caused by an APC mutation. *Human Genome Variation*. 2: 15011, 2015.
15. Inazawa J, Muramatsu T, Kozaki K, Imoto S, Yamaguchi R, Tsuda H, Kawano T, Fujiwara N, Morishita M, Miyano S. The hypusine cascade promotes cancer progression and metastasis through the regulation of RhoA in squamous cell carcinoma. *Oncogene*. In press.
16. Ito S, Shiraishi Y, Shimamura T, Chiba K, Miyano S. High performance computing of a fusion gene detection pipeline on the K computer.

BIBM 2015. 1441-1447, 2015.
17. Ito H, Shiwaku H, Yoshida C, Homma H, Luo H, Chen X, Fujita K, Musante L, Fischer U, Frints SG, Romano C, Ikeuchi Y, Shimamura T, Imoto S, Miyano S, Muramatsu SI, Kawauchi T, Hoshino M, Sudol M, Arumughan A, Wanker EE, Rich T, Schwartz C, Matsuzaki F, Bonni A, Kalscheuer VM, Okazawa H. In utero gene therapy rescues microcephaly caused by Pqbp1-hypofunction in neural stem progenitor cells. *Mol Psychiatry*. 20(4): 459-471, 2015.
18. Iwakawa R, Kohno T, Totoki Y, Shibata T, Tsuchihara K, Mimaki S, Tsuta K, Narita Y, Nishikawa R, Noguchi M, Harris CC, Robles AI, Yamaguchi R, Imoto S, Miyano S, Totsuka H, Yoshida T, Yokota J. Expression and clinical significance of genes frequently mutated in small cell lung cancers defined by whole exome/RNA sequencing. *Carcinogenesis*. 36(6): 616-621, 2015.
19. Kanojia D, Nagata Y, Garg M, Lee DH, Sato A, Yoshida K, Sato Y, Sanada M, Mayakonda A, Bartenhagen C, Klein HU, Doan NB, Said JW, Mohith S, Gunasekar S, Shiraishi Y, Chiba K, Tanaka H, Miyano S, Myklebost O, Yang H, Dugas M, Meza-Zepeda LA, Silberman AW, Forscher C, Tyner JW, Ogawa S, Koeffler HP. Genomic landscape of liposarcoma. *Oncotarget*. 6(40): 42429-42444, 2015.
20. Kataoka K, Nagata Y, Kitanaka A, Shiraishi Y, Shimamura T, Yasunaga J, Totoki Y, Chiba K, Sato-Otsubo A, Nagae G, Ishii R, Muto S, Kotani S, Watatani Y, Takeda J, Sanada M, Tanaka H, Suzuki H, Sato Y, Shiozawa Y, Yoshizato T, Yoshida K, Makishima H, Iwanaga M, Ma G, Nosaka K, Hishizawa M, Itonaga H, Imaizumi Y, Munakata W, Ogasawara H, Sato T, Sasai K, Muramoto K, Penova M, Kawaguchi T, Nakamura H, Hama N, Shide K, Kubuki Y, Hidaka T, Kameda T, Nakamaki T, Ishiyama K, Miyawaki S, Yoon SS, Tobinai K, Miyazaki Y, Takaori-Kondo A, Matsuda F, Takeuchi K, Nureki O, Aburatani H, Watanabe T, Shibata T, Matsuoka M, Miyano S, Shimoda K, Ogawa S. Integrated molecular analysis of adult T cell leukemia/lymphoma. *Nat Genet*. 47(11): 1304-1315, 2015.
21. Kato M, Seki M, Yoshida K, Sato Y, Oyama R, Arakawa Y, Kishimoto H, Taki T, Akiyama M, Shiraishi Y, Chiba K, Tanaka H, Mitsuiki N, Kajiwara M, Mizutani S, Sanada M, Miyano S, Ogawa S, Koh K, Takita J. Genomic analysis of clonal origin of Langerhans cell histiocytosis following acute lymphoblastic leukaemia. *Br J Haematol*. 2015 Nov 5. doi: 10.1111/bjh.13841. [Epub ahead of print]
22. Kawashima-Goto S, Imamura T, Seki M, Kato M, Yoshida K, Sugimoto A, Kaneda D, Fujiki A, Miyachi M, Nakatani T, Osone S, Ishida H, Taki T, Takita J, Shiraishi Y, Chiba K, Tanaka H, Miyano S, Ogawa S, Hosoi H. Identification of a homozygous JAK3 V674A mutation caused by acquired uniparental disomy in a relapsed early T-cell precursor ALL patient. *Int J Hematol*. 101 (4): 411-416, 2015.
23. Kayano M, Matsui H, Yamaguchi R, Imoto S, Miyano S. Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to time-dependent biomarker detection. *Biostatistics*. 2015 Sep 28. pii: kxv037. [Epub ahead of print]
24. Koso H, Yi Hungtsung Y, Sheridan P, Miyano S, Ino Y, Todo T, Watanabe S. Identification of LARP4B as a Tumor Suppressor Gene in Glioma. *Can Res*. In press.
25. Kurtovic-Kozaric A, Przychodzen B, Singh J, Konarska MM, Clemente MJ, Otrock ZK, Nakashima M, Hsi ED, Yoshida K, Shiraishi Y, Chiba K, Tanaka H, Miyano S, Ogawa S, Boultwood J, Makishima H, Maciejewski JP, Padgett RA. PRPF8 defects cause missplicing in myeloid malignancies. *Leukemia*. 29(1): 126-136, 2015.
26. Li Y, Shibuya T. Malphite: A convolutional neural network and ensemble learning based protein secondary structure predictor. *BIBM 2015*. 1260-1266, 2015.
27. Madan V, Kanojia D, Li J, Okamoto R, Sato-Otsubo A, Kohlmann A, Sanada M, Grossmann V, Sundaresan J, Shiraishi Y, Miyano S, Thol F, Ganser A, Yang H, Haferlach T, Ogawa S, Koeffler HP. Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. *Nat Commun*. 6: 6042, 2015.
28. Manolio TA, Abramowicz M, Al-Mulla F, Anderson W, Balling R, Berger AC, Bleyl S, Chakravarti A, Chantratita W, Chisholm RL, Dissanayake VH, Dunn M, Dzau VJ, Han BG, Hubbard T, Kolbe A, Korf B, Kubo M, Lasko P, Leego E, Mahasirimongkol S, Majumdar PP, Matthijs G, McLeod HL, Metspalu A, Meulien P, Miyano S, Naparstek Y, O'Rourke PP, Patrinos GP, Rehm HL, Relling MV, Rennert G, Rodriguez LL, Roden DM, Shuldiner AR, Sinha S, Tan P, Ulfendahl M, Ward R, Williams MS, Wong JE, Green ED, Ginsburg GS. Global implementation of genomic medicine: We are not alone. *Sci Transl Med*. 7(290): 290ps13, 2015.
29. Molenaar RJ, Thota S, Nagata Y, Patel B, Clemente M, Przychodzen B, Hirsh C, Viny AD, Hosano N, Bleeker FE, Meggendorfer M, Alpermann T, Shiraishi Y, Chiba K, Tanaka H, van Noorden CJ, Radivoyevitch T, Carraway HE, Makishima H, Miyano S, Sekeres MA, Ogawa S, Haferlach T, Maciejewski JP. Clinical and biological implications of ancestral and non-ancestral IDH1 and IDH2 mutations in myeloid neoplasms. *Leukemia*. 29(11): 2134-2142,

2015.
30. Nagata Y, Kontani K, Enami T, Kataoka K, Ishii R, Totoki Y, Kataoka TR, Hirata M, Aoki K, Nakano K, Kitanaka A, Sakata-Yanagimoto M, Egami S, Shiraishi Y, Chiba K, Tanaka H, Shiozawa Y, Yoshizato T, Suzuki H, Kon A, Yoshida K, Sato Y, Sato-Otsubo A, Sanada M, Munakata W, Nakamura H, Hama N, Miyano S, Nureki O, Shibata T, Haga H, Shimoda K, Katada T, Chiba S, Watanabe T, Ogawa S. Variegated RHOA mutations in adult T-cell leukemia/lymphoma. *Blood*. 2015 Nov 16. pii:blood-2015-06-644948. [Epub ahead of print]
31. Nakata A, Yoshida R, Yamaguchi R, Yamauchi M, Tamada Y, Fujita A, Shimamura T, Imoto S, Higuchi T, Nomura M, Kimura T, Nokihara H, Higashiyama M, Kondoh K, Nishihara H, Tojo A, Yano S, Miyano S, Gotoh N. Elevated β-catenin pathway as a novel target for patients with resistance to EGF receptor targeting drugs. *Sci Rep*. 5: 13076, 2015.
32. Okuno Y, Hoshino A, Muramatsu H, Kawashima N, Wang X, Yoshida K, Wada T, Gunji M, Toma T, Kato T, Shiraishi Y, Iwata A, Hori T, Kitoh T, Chiba K, Tanaka H, Sanada M, Takahashi Y, Nonoyama S, Ito M, Miyano S, Ogawa S, Kojima S, Kanegane H. Late-Onset Combined Immunodeficiency with a Novel IL2 RG Mutation and Probable Revertant Somatic Mosaicism. *J Clin Immunol*. 35(7): 610-614, 2015.
33. Park H, Niida A, Miyano S, Imoto S. Sparse overlapping group lasso for integrative multi-omics analysis. *J Comput Biol*. 22(2): 73-84, 2015.
34. Park H, Imoto S, Miyano S. Recursive Random Lasso (RRLasso) for Identifying Anti-Cancer Drug Targets. *PLoS One*. 10(11): e0141869, 2015.
35. Polprasert C, Schulze I, Sekeres MA, Makishima H, Przychodzen B, Hosono N, Singh J, Padgett RA, Gu X, Phillips JG, Clemente M, Parker Y, Lindner D, Dienes B, Jankowsky E, Sauntharajah Y, Du Y, Oakley K, Nguyen N, Mukherjee S, Pabst C, Godley LA, Churpek JE, Pollyea DA, Krug U, Berdel WE, Klein HU, Dugas M, Shiraishi Y, Chiba K, Tanaka H, Miyano S, Yoshida K, Ogawa S, Müller-Tidow C, Maciejewski JP. Inherited and Somatic Defects in DDX41 in Myeloid Neoplasms. *Cancer Cell*. 27 (5): 658-670, 2015.
36. Saini H, Raicar G, Sharma A, Lal S, Dehzangi A, Lyons J, Paliwal KK, Imoto S, Miyano S. Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition. *J Theor Biol*. 380: 291-298, 2015.
37. Sakaguchi H, Muramatsu H, Okuno Y, Makishima H, Xu Y, Furukawa-Hibi Y, Wang X, Narita A, Yoshida K, Shiraishi Y, Doisaki S, Yoshida N, Hama A, Takahashi Y, Yamada K, Miyano S, Ogawa S, Maciejewski JP, Kojima S. Aberrant DNA Methylation Is Associated with a Poor Outcome in Juvenile Myelomonocytic Leukemia. *PLoS One*. 10(12): e0145394, 2015.
38. Sasaki Y, Shibuya T, Ito K, Arimura H. Efficient Approximate 3-Dimensional Point Set Matching Using Root-Mean-Square Deviation Score. Lecture Notes in Computer Science. 9371: 191-203, 2015.
39. Sawada G, Niida A, Hirata H, Komatsu H, Uchi R, Shimamura T, Takahashi Y, Kurashige J, Matsumura T, Ueo H, Takano Y, Ueda M, Sakimura S, Shinden Y, Eguchi H, Sudo T, Sugimachi K, Yamasaki M, Tanaka F, Tachimori Y, Kajiyama Y, Natsugoe S, Fujita H, Tanaka Y, Calin G, Miyano S, Doki Y, Mori M, Mimori K. An Integrative Analysis to Identify Driver Genes in Esophageal Squamous Cell Carcinoma. *PLoS One*. 10(10): e0139808, 2015.
40. Seki M, Nishimura R, Yoshida K, Shimamura T, Shiraishi Y, Sato Y, Kato M, Chiba K, Tanaka H, Hoshino N, Nagae G, Shiozawa Y, Okuno Y, Hosoi H, Tanaka Y, Okita H, Miyachi M, Souzaki R, Taguchi T, Koh K, Hanada R, Kato K, Nomura Y, Akiyama M, Oka A, Igarashi T, Miyano S, Aburatani H, Hayashi Y, Ogawa S, Takita J. Integrated genetic and epigenetic analysis defines novel molecular subgroups in rhabdomyosarcoma. *Nat Commun*. 6: 7557, 2015.
41. Shiota M, Yang X, Kubokawa M, Morishima T, Tanaka K, Mikami M, Yoshida K, Kikuchi M, Izawa K, Nishikomori R, Okuno Y, Wang X, Sakaguchi H, Muramatsu H, Kojima S, Miyano S, Ogawa S, Takagi M, Hata D, Kanegane H. Somatic mosaicism for a NRAS mutation associates with disparate clinical features in RAS-associated leukoproliferative disease: a report of two cases. *J Clin Immunol*. 35(5): 454-8, 2015.
42. Shiraishi Y, Tremmel G, Miyano S, Stephens M. A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. *PLoS Genet*. 11(12): e1005657, 2015.
43. Sung W-K, Sadakane K, Shibuya T, Belorkar A, Pyrogova I. An O(m, log m)-Time Algorithm for Detecting Superbubbles. *IEEE/ACM Trans Comput Biology Bioinform*. 12(4): 770-777, 2015.
44. Suzuki H, Aoki K, Chiba K, Sato Y, Shiozawa Y, Shiraishi Y, Shimamura T, Niida A, Motomura K, Ohka F, Yamamoto T, Tanahashi K, Ranjit M, Wakabayashi T, Yoshizato T, Kataoka K, Yoshida K, Nagata Y, Sato-Otsubo A, Tanaka H, Sanada M, Kondo Y, Nakamura H, Mizoguchi M, Abe T, Muragaki Y, Watanabe R, Ito I, Miyano S, Natsume A, Ogawa S. Mutational landscape and clonal architecture in grade II and III gliomas. *Nat Genet*. 47(5): 458-468, 2015.
45. Tagawa K, Homma H, Saito A, Fujita K, Chen X, Imoto S, Oka T, Ito H, Motoki K, Yoshida C, Hatsuta H, Murayama S, Iwatsubo T, Miyano S,

Okazawa H. Comprehensive phosphoproteome analysis unravels the core signaling network that initiates the earliest synapse pathology in preclinical Alzheimer's disease brain. *Hum Mol Genet*. 4(2): 540-58, 2015.

46. Takahashi Y, Sheridan P, Niida A, Sawada G, Uchi R, Mizuno H, Kurashige J, Sugimachi K, Sasaki S, Shimada Y, Hase K, Kusunoki M, Kudo S, Watanabe M, Yamada K, Sugihara K, Yamamoto H, Suzuki A, Doki Y, Miyano S, Mori M, Mimori K. The AURKA/TPX2 axis drives colon tumorigenesis cooperatively with MYC. *Ann Oncol*. 26(5): 935-42, 2015.
47. Tamura K, Hazama S, Yamaguchi R, Imoto S, Takenouchi H, Inoue Y, Kanekiyo S, Shindo Y, Satoru Miyano S, Nakamura Y, Kiyotani K, Characterization of T cell repertoire in tumor tissues and blood in advanced colorectal cancers through deep T cell receptor sequencing. *Oncology Letters*. In press.
48. Wang R, Yoshida K, Toki T, Sawada T, Uechi T, Okuno Y, Sato-Otsubo A, Kudo K, Kamimaki I, Kanezaki R, Shiraishi Y, Chiba K, Tanaka H, Terui K, Sato T, Iribe Y, Ohga S, Kuramitsu M, Hamaguchi I, Ohara A, Hara J, Goi K, Matsubara K, Koike K, Ishiguro A, Okamoto Y, Watanabe K, Kanno H, Kojima S, Miyano S, Kenmochi N, Ogawa S, Ito E. Loss of function mutations in RPL27 and RPS27 identified by whole-exome sequencing in Diamond-Blackfan anaemia. *Br J Haematol*. 168(6): 854-64, 2015.
49. Wang X, Muramatsu H, Okuno Y, Sakaguchi H, Yoshida K, Kawashima N, Xu Y, Shiraishi Y, Chiba K, Tanaka H, Saito S, Nakazawa Y, Masunari T, Hirose T, Elmahdi S, Narita A, Doisaki S, Ismael O, Makishima H, Hama A, Miyano S, Takahashi Y, Ogawa S, Kojima S. GATA2 and secondary mutations in familial myelodysplastic syndromes and pediatric myeloid malignancies. *Haematologica*. 100(10): e 398-401, 2015.
50. Yamaguchi H, Sakaguchi H, Yoshida K, Yabe M, Yabe H, Okuno Y, Muramatsu H, Takahashi Y, Yui S, Shiraishi Y, Chiba K, Tanaka H, Miyano S, Inokuchi K, Ito E, Ogawa S, Kojima S. Clinical and genetic features of dyskeratosis congenita, cryptic dyskeratosis congenita, and Hoyeraal-Hreidarsson syndrome in Japan. *Int J Hematol*. 102(5): 544-552, 2015.
51. Yamaguchi R, Imoto S, Miyano S. A TCR sequence data analysis pipeline: Tcrip. In *Immunopharmacogenomics*. Nakamura, Y. (Ed.), Springer Japan, pp. 27-43, 2015.
52. Yamaguchi K, Komura M, Yamaguchi R, Imoto S, Shimizu E, Kasuya S, Shibuya T, Hatakeyama S, Takahashi N, Ikenoue T, Hata K, Tsurita G, Shinozaki M, Suzuki Y, Sugano S, Miyano S, Furukawa Y. Detection of APC mosaicism by next-generation sequencing in an FAP patient. *J Hum Genet*. 60(5): 227-231, 2015.
53. Yang X, Hoshino A, Taga T, Kunitsu T, Ikeda Y, Yasumi T, Yoshida K, Wada T, Miyake K, Kubota T, Okuno Y, Muramatsu H, Adachi Y, Miyano S, Ogawa S, Kojima S, Kanegane H. A female patient with incomplete hemophagocytic lymphohistiocytosis caused by a heterozygous XIAP mutation associated with non-random X-chromosome inactivation skewed towards the wild-type XIAP allele. *J Clin Immunol*. 35(3): 244-248, 2015.
54. Yasuda T, Miyano S. Inferring the global structure of chromosomes from structural variations. *BMC Genomics*. 16 Suppl 2: S13, 2015. Yasuda T, Miyano S. Erratum to: inferring the global structure of chromosomes from structural variations. *BMC Genomics*. 16: 276, 2015.
55. Yew PY, Alachkar H, Yamaguchi R, Kiyotani K, Fang H, Yap KL, Liu HT, Wickrema A, Artz A, van Besien K, Imoto S, Miyano S, Bishop MR, Stock W, Nakamura Y. Quantitative characterization of T-cell repertoire in allogeneic hematopoietic stem cell transplant recipients. *Bone Marrow Transplant*. 50(9): 1227-34, 2015.
56. Yoshizato T, Dumitriu B, Hosokawa K, Makishima H, Yoshida K, Townsley D, Sato-Otsubo A, Sato Y, Liu D, Suzuki H, Wu CO, Shiraishi Y, Clemente MJ, Kataoka K, Shiozawa Y, Okuno Y, Chiba K, Tanaka H, Nagata Y, Katagiri T, Kon A, Sanada M, Scheinberg P, Miyano S, Maciejewski JP, Nakao S, Young NS, Ogawa S. Somatic Mutations and Clonal Hematopoiesis in Aplastic Anemia. *N Engl J Med*. 373 (1): 35-47, 2015.

Human Genome Center

Laboratory of Molecular Medicine
ゲノム医科学分野

Professor	Tatsuhiro Shibata, M.D., Ph.D.	教　授	柴　田　龍　弘
Assistant Professor	Satoshi Yamasaki, Ph.D.	助　教	山　﨑　　智

The Laboratory of Molecular Medicine focuses on comprehensive characterization of currently-untreatable diseases including cancer on the basis of molecular genomics and aims to make "breakthroughs for human health" by identifying novel disease-related genes/pathways, including potential therapeutic or preventive targets and biomarkers, and to understand human diseases as heterogeneous but intervention-able "biological systems". This group has also organized the facility for the analysis of next-generation high-performance sequencers.

1. Comprehensive molecular analysis of biliary tract cancer

Hiromi Nakamura[1], Yasuhito Arai[1], Yasushi Totoki[1], Tomoki Shirota[1,2], Asmaa Elzawahry[1], Mamoru Kato[3], Natsuko Hama[1], Fumie Hosoda[1], Tomoko Urushidate[4], Shoko Ohashi[1], Nobuyoshi Hiraoka[5], Hidenori Ojima[5,6], Kazuaki Shimada[7], Takuji Okusaka[8], Tomoo Kosuge[7], Shinichi Miyagawa[2], Tatsuhiro Shibata[1,4,9]: [1]Division of Cancer Genomics, [3]Department of Bioinformatics, National Cancer Center Research Institute, [2]First Department of Surgery, Shinshu University School of Medicine, [4]Laboratory of Molecular Medicine, Human Genome Center, The institute of Medical Science, the University of Tokyo, [5]Division of Pathology and Clinical Laboratories, [7]Hepatobiliary and Pancreatic Surgery Division, [8]Department of Hepatobiliary and Pancreatic Oncology, National Cancer Center Hospital, [6]Department of Pathology, Keio University School of Medicine.

The incidence of biliary tract cancer, including intra-hepatic (ICC) and extra-hepatic (ECC) cholangiocarcinomas and gallbladder cancer (GB), has rapidly increased globally; however, no effective targeted molecular therapies have been approved and the cure rate is very low. This study characterized more than 200 BTCs by a combination of exome and transcriptome sequencing and uncovered spectrums of molecular alterations that included novel therapeutic targets. Gradient spectrum of mutational signatures with higher burden of the APOBEC-associated signature in GB and ECC was observed. Thirty-two driver genes were identified as significantly altered genes and nearly 40% of cases harbored molecularly targetable genetic alterations. Organ-specific dysfunctions were also identified in epigenetic regulators, often in unique combinations with growth-promoting alterations (e.g., IDH1 mutations without RAS mutations). The group with the poorest prognosis had a characteristic elevation of both immune response signatures and counteracting immune checkpoint molecules, implying that immune-modulating therapies could also be potentially promising options for those patients.

2. Pathway and network analysis of cancer genomes

Pau Creixell[1], Jüri Reimand[2], Syed Haider[3], Guanming Wu[3,4], Tatsuhiro Shibata[5], Miguel Vazquez[6], Ville Mustonen[7], Abel Gonzalez-Perez[8], John Pearson[9], Chris Sander[10], Benjamin J. Raphael[11], Debora Marks[12], B.F. Francis Ouellette[3,20], Alfonso

Valencia[6], Gary D. Bader[2], Paul C. Boutros[3,13,14], Joshua M. Stuart[15,16], Rune Linding[1,17], Nuria Lopez-Bigas[8,18], Lincoln D. Stein[3,19]: [1]Cellular Signal Integration Group (C-SIG), Technical University of Denmark, Lyngby, Denmark, [2]The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada, [3]Informatics and Biocomputing Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada, [4]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA, [5]Laboratory of Molecular Medicine, Human Genome Center, The institute of Medical Science, the University of Tokyo, [6]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre, Madrid, Spain, [7]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, [8]Research Unit on Biomedical Informatics, University Pompeu Fabra, Barcelona, Spain. [9]Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, University of Queensland, St. Lucia, Brisbane, Queensland, Australia, [10]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, NY, USA, [11]Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, RI, USA, [12]Department of Systems Biology, Harvard Medical School, Boston, MA USA, [13]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada, [14]Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada, [15]Department of Biomolecular Engineering, University of California, Santa Cruz, California, USA, [16]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California, USA, [17]Biotech Research & Innovation Centre (BRIC), University of Copenhagen (UCPH), DK-2200 Copenhagen, Denmark, [18]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain, [19]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada, [20]Department of Cell and Systems Biology, University of Toronto, Toronto, Ontario, Canada.

Genomic information on tumors from 50 cancer types catalogued by The International Cancer Genome Consortium (ICGC) shows that only few well-studied driver genes are frequently mutated, in contrast to many infrequently mutated genes that may also contribute to tumor biology. This international collaborative study provides an overview of these pathway analysis techniques to guide mechanistic and translational investigations.

Publications

1. Nakamura H, Arai Y, Totoki Y, Shirota T, Elzawahry A, Kato M, Hama N, Hosoda F, Urushidate T, Ohashi S, Hiraoka N, Ojima H, Shimada K, Okusaka T, Kosuge T, Miyagawa S, Shibata T. Genomic spectra of biliary tract cancer. Nat Genet, 2015, 47: 1003-10.
2. Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, Mustonen V, Gonzalez-Perez A, Pearson J, Sander C, Raphael1 BJ, Marks1 DS, Ouellette BF, Valencia A, Bader GD, Boutros PC, Stuart JM, Linding R, Lopez-Bigas N, Stein LD. Pathway and network analysis of cancer genomes. Nat Methods, 2015 12: 615-21.
3. Kataoka K, Nagata Y, Kitanaka A, Shiraishi Y, Shimamura T, Yasunaga JI, Totoki Y, Chiba K, Sato-Otsubo A, Nagae G, Ishii R, Muto S, Kotani S, Watatani Y, Takeda J, Sanada M, Tanaka H, Suzuki H, Sato Y, Shiozawa Y, Yoshizato T, Yoshida K, Makishima H, Iwanaga M, Ma G, Nosaka K, Hishizawa M, Itonaga H, Imaizumi Y, Munakata W, Ogasawara H, Sato T, Sasai K, Muramoto K, Penova M, Kawaguchi T, Nakamura H, Hama N, Shide K, Kubuki Y, Hidaka T, Kameda T, Nakamaki T, Ishiyama K, Miyawaki S, Yoon SS, Tobinai K, Miyazaki Y, Takaori-Kondo A, Matsuda F, Takeuchi K, Nureki O, Aburatani H, Watanabe T, Shibata T, Matsuoka M, Miyano S, Shimoda K, Ogawa S. Integrated molecular analysis of adult T cell leukemia/lymphoma. Nat Genet. 2015, 47: 1304-15.

Human Genome Center

Laboratory of Genome Technology
シークエンス技術開発分野

Professor	Satoru Miyano, Ph.D.	教　授	宮　野　　悟(兼)
Associate Professor	Koichi Matsuda, M.D., Ph.D.	准教授	松　田　浩　一
Assistant Professor	Chizu Tanikawa, Ph.D.	助　教	谷　川　千　津
Project Assistant Professor	Takafumi Miyamoto, Ph.D.	特任助教	宮　本　崇　史
Project Assistant Professor	Makoto Hirata, M.D., Ph.D.	特任助教	平　田　　真

The major goal of our group is to identify genes 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 RNA sequence method, we have isolated a number of biologically and/or medically important genes, and are developing novel diagnostic and therapeutic tools.

Koichi Matsuda, Chizu Tanikawa, Takafumi Miyamoto, Makoto Hirata, Jinichi Mori, Tomoyuki Koguchi, Yusuke Tsuda, Ryuta Yamamoto, Varalee Yodsurang, Yukie Takahashi, Wang Guanxiong, Satomi Takahashi, Misato Oshima.

1. Genome-wide association study

Genome-wide association study identified SNP on 15q24 associated with bladder cancer risk in Japanese population.

Through genome-wide association analysis and an independent replication study using a total of 1131 bladder cancer cases and 12 558 non-cancer controls of Japanese populations, we identified a susceptibility locus on chromosome 15q24. SNP rs11543198 was associated with bladder cancer risk with odds ratio (OR) of 1.41 and P-value of 4.03×10^{-9}. Subgroup analysis revealed rs11543198 to have a stronger effect in male smokers with OR of 1.66. SNP rs8041357, which is in complete linkage disequilibrium ($r^2 = 1$) with rs11543198, was also associated with bladder cancer risk in Europeans ($P = 0.045$ for an additive and $P = 0.025$ for a recessive model), despite much lower minor allele frequency in Europeans (3.7%) compared with the Japanese (22.2%). Imputational analysis in this region suggested CYP1A2, which metabolizes tobacco-derived carcinogen, as a causative candidate gene. We also confirmed the association of previously reported loci, namely SLC14A1, APOBEC3A, PSCA and MYC, with bladder cancer. Our finding implies the crucial roles of genetic variations on the chemically associated development of bladder cancer.

Large-scale association analysis in Asians identifies new susceptibility loci for prostate cancer.

Genome-wide association studies (GWAS) have identified ～100 genetic loci associated with prostate cancer risk. Less than a dozen of these loci were initially identified from GWAS in two Asian populations, likely because of smaller sample sizes of these individual GWAS in Asians. Here, we conduct a large-scale meta-analysis of two GWAS from the Japanese population (1,583 cases and 3,386 controls) and the Chinese population (1,417 cases and 1,008 controls), followed by replication in three independent sample sets. We identify two independ-

ent susceptibility loci for prostate cancer at 11p15.4 (rs12791447, $P = 3.59 \times 10^{-8}$; PPFIBP2) and 14q23.2 (rs58262369, $P = 6.05 \times 10^{-10}$; ESR2). The mRNA levels of PPFIBP2 and ESR2 are differentially expressed in prostate tumours and paired normal tissues. Our study adds two new loci to the limited number of prostate cancer risk-associated variants in Asians and provides important insight into potential biological mechanisms of prostate cancer.

Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese.

To fine map association signals of human leukocyte antigen (HLA) variants in the major histocompatibility complex (MHC) region, we constructed a Japanese population-specific reference panel (n = 908). We conducted trans-ancestry comparisons of linkage disequilibrium (LD) and haplotype structure for HLA variants using an entropy-based LD measurement, ε, and a visualization tool to capture high-dimensional variables. Our Japanese reference panel exhibited stronger LD between HLA genes than European or other East Asian populations, characterized by one population-specific common long-range HLA haplotype. We applied HLA imputation to genome-wide association study (GWAS) data for Graves' disease in Japanese (n = 9,003) and found that amino acid polymorphisms of multiple class I and class II HLA genes independently contribute to disease risk (HLA-DPB1, HLA-A, HLA-B and HLA-DRB1; $P < 2.3 \times 10^{-6}$), with the strongest impact at HLA-DPB1 ($P = 1.6 \times 10^{-42}$). Our study illustrates the value of population-specific HLA reference panels.

A genome-wide association study identifies PLCL2 and AP3D1-DOT1L-SF3A2 as new susceptibility loci for myocardial infarction in Japanese.

Despite considerable progress in preventive and therapeutic strategies, myocardial infarction (MI) is one of the leading causes of death throughout the world. A total of 55 susceptibility genes have been identified mostly in European genome-wide association studies (GWAS). Nevertheless, large-scale GWAS from other population could possibly find additional susceptibility loci. To identify as many MI susceptibility loci as possible, we performed a large-scale genomic analysis in Japanese population. To identify MI susceptibility loci in Japanese, we conducted a GWAS using 1666 cases and 3198 controls using the Illumina Human610-Quad BeadChip and HumanHap550v3 Genotyping BeadChip. We performed replication studies using a total of 11,412 cases and 28,397 controls in the Japanese population. Our study identified two novel susceptibility loci for MI: PLCL2 on chromosome 3p24.3 (rs4618210: A > G, $P = 2.60 \times 10^{-9}$, odds ratio (OR) = 0.91) and AP3D1-DOT1L-SF3A2 on chromosome 19 p13.3 (rs3803915: A > C, $P = 3.84 \times 10^{-9}$, OR = 0.89). Besides, a total of 14 previously reported MI susceptibility loci were replicated in our study. In particular, we validated a strong association on chromosome 12q24 (rs3782886: A > G: $P = 1.14 \times 10^{-14}$, OR = 1.46). Following pathway analysis using 265 genes related to MI or coronary artery disease, we found that these loci might be involved in the pathogenesis of MI via the promotion of atherosclerosis. In the present large-scale genomic analysis, we identified PLCL2 and AP3D1-DOT1L-SF3A2 as new susceptibility loci for MI in the Japanese population. Our findings will add novel findings for MI susceptibility loci.

2. Genes playing significant roles in human cancers

(1) p53 pathway

Regulation of iron homeostasis by the p53-ISCU pathway.

Accumulation of iron in tissues increases the risk of cancer, but iron regulatory mechanisms in cancer tissues are largely unknown. Here, we report that p53 regulates iron metabolism through the transcriptional regulation of ISCU (iron-sulfur cluster assembly enzyme), which encodes a scaffold protein that plays a critical role in Fe-S cluster biogenesis. p53 activation induced ISCU expression through binding to an intronic p53-binding site. Knockdown of ISCU enhanced the binding of iron regulatory protein 1 (IRP1), a cytosolic Fe-S protein, to an iron-responsive element in the 5' UTR of ferritin heavy polypeptide 1 (FTH1) mRNA and subsequently reduced the translation of FTH1, a major iron storage protein. In addition, in response to DNA damage, p53 induced FTH1 and suppressed transferrin receptor, which regulates iron entry into cells. HCT116 $p53^{+/+}$ cells were resistant to iron accumulation, but HCT116 $p53^{-/-}$ cells accumulated intracellular iron after DNA damage. Moreover, excess dietary iron caused significant elevation of serum iron levels in $p53^{-/-}$ mice. ISCU expression was decreased in the majority of human liver cancer tissues, and its reduced expression was significantly associated with p53 mutation. Our finding revealed a novel role of the p53-ISCU pathway in the maintenance of iron homeostasis in hepatocellular carcinogenesis.

Cystatin C as a p53-inducible apoptotic mediator which regulates Cathepsin L activity.

In response to various cellular stresses, p53 is ac-

tivated and inhibits malignant transformation through the transcriptional regulation of its target genes. However, the full picture of the p53 downstream pathway still remains to be elucidated. Here we identified cystatin C, a major inhibitor of cathepsins, as a novel p53-target. In response to DNA damage, activated p53 induced cystatin C expression through p53 binding sequence in the first intron. We showed that cathepsin L activity was decreased in HCT116 $p53^{+/+}$ cells after adriamycin treatment, but not in HCT116 $p53^{-/-}$ cells. We also found that knockdown of cystatin C reduced adriamycin-induced caspase-3 activation. Cystatin C expression was significantly downregulated in breast cancer cells with p53 mutations, and decreased cystatin C expression was associated with poor prognosis of breast cancer. Our findings revealed an important role of p53-cystatin C pathway in human carcinogenesis.

(2) Breast cancer

Identification of novel epigenetically inactivated gene PAMR1 in breast carcinoma.

Development of cancer is a complex process involving multiple genetic and epigenetic alterations. In our microarray analysis of 81 breast carcinoma specimens, we identified peptidase domain containing associated with muscle regeneration 1 (PAMR1) as being frequently suppressed in breast cancer tissues. PAMR1 expression was also reduced in all tested breast cancer cell lines, while PAMR1 was expressed moderately in normal breast tissues and primary mammary epithelial cells. DNA sequencing of the PAMR1 promoter after sodium bisulfite treatment revealed that CpG sites were hypermethylated in the breast cancer tissues and cell lines. PAMR1 expression was restored by 5-aza-2' deoxycytidine treatment, demonstrating that promoter hypermethylation contributed to PAMR1 inactivation in the breast cancer cells. In addition, ectopic expression of PAMR1 markedly suppressed cancer cell growth. In summary, our study identified PAMR1 as a putative tumor suppressor which was frequently inactivated by promoter hypermethylation in breast cancer tissues.

Publications

[1] Mori J, Tanikawa C, Funauchi Y, Lo PH, Nakamura Y, and Matsuda K. Cystatin C as a p53-inducible apoptotic mediator which regulates Cathepsin L activity. Cancer Science. In Press.

[2] Funauchi Y, Tanikawa C, Yi Lo PH, Mori J, Daigo Y, Takano A, Miyagi Y, Okawa A, Nakamura Y, and Matsuda K. Regulation of iron homeostasis by the p53-ISCU pathway. Scientific reports. 5: 16497, 2015.

[3] Wang M, Takahashi A, Liu F, Ye D, Ding Q, Qin C, Yin C, Zhang Z, Matsuda K, Kubo M, Na R, Lin X, Jiang H, Ren S, Sun J, Zheng SL, Le Marchand L, Isaacs WB, Mo Z, Haiman CA, Sun Y, Nakagawa H, and Xu J. Large-scale association analysis in Asians identifies new susceptibility loci for prostate cancer. Nature communications. 6: 8469, 2015.

[4] Joshi PK, Esko T, Matsuda K, Wilson JF, et al. Directional dominance on stature and cognition in diverse human populations, Nature, 523: 459-462, 2015.

[5] Kim YK, Oh JH, Kim YJ, Hwang MY, Moon S, Low SK, Takahashi A, Matsuda K, Kubo M, Lee J, and Kim BJ. Influence of Genetic Variants in EGF and Other Genes on Hematological Traits in Korean Populations by a Genome-Wide Approach. Biomed Res Int, 2015: 914965, 2015.

[6] Okada Y, Momozawa Y, Ashikawa K, Kanai M, Matsuda K, Kamatani Y, Takahashi A, and Kubo M. Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. Nat Genet. 47: 798-802, 2015.

[7] Fang H, Yamaguchi R, Liu X, Daigo Y, Yew PY, Tanikawa C, Matsuda K, Imoto S, Miyano S, and Nakamura Y. Quantitative T cell repertoire analysis by deep cDNA sequencing of T cell receptor α and β chains using next-generation sequencing (NGS). Oncoimmunology. 3 (12): e968467, 2015.

[8] Lo PH, Tanikawa C, Katagiri T, Nakamura Y, and Matsuda K. Identification of novel epigenetically inactivated gene PAMR1 in breast carcinoma, Oncol Rep. 33(1): 267-273, 2015.

[9] Matsuda K, Takahashi A, Middlebrooks CD, Obara W, Nasu Y, Inoue K, Tamura K, Yamasaki I, Naya Y, Tanikawa C, Cui R, Figueroa JD, Silverman DT, Rothman N, Namiki M, Tomita Y, Nishiyama H, Kohri K, Deguchi T, Nakagawa M, Yokoyama M, Miki T, Kumon H, Fujioka T, Prokunina-Olsson L, Kubo M, Nakamura Y, and Shuin T. Genome-wide association study identified SNP on 15q24 associated with bladder cancer risk in Japanese population. Hum Mol Genet. 24(4): 1177-1184, 2015.

[10] Hirokawa M, Morita H, Tajima T, Takahashi A, Ashikawa K, Miya F, Shigemizu D, Ozaki K, Sakata Y, Nakatani D, Suna S, Imai Y, Tanaka T, Tsunoda T, Matsuda K, Kadowaki

T, Nakamura Y, Nagai R, Komuro I, and Kubo M. A genome-wide association study identifies PLCL2 and AP3D1-DOT1L-SF3A2 as new susceptibility loci for myocardial infarction in Japanese. Eur J Hum Genet. 23(3): 374-380, 2015.

[11] 谷川千津，松田浩一．がんとシトルリン化タンパク質．医学のあゆみ255(9)：871−876，2015.

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	博士(理学)　中　井　謙　太
Senior Assistant Professor	Ashwini Patil, Ph.D.		講　師	博士(理学)　パティル，アシュウィニ
Project Senior Assistant Professor	Sung-Joon Park, Ph.D.		特任講師	博士(工学)　朴　　聖　俊

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 under which 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 interactions.

1. Genome-wide identification and characterization of transcription start sites and promoters in the tunicate *Ciona intestinalis*

Rui Yokomori, Kotaro Shimai[1], Koki Nishitsuji[2], Yutaka Suzuki[3], Takehiro G. Kusakabe[1] and Kenta Nakai: [1]Fac. of Sci. and Engineering, Konan Univ, [2]Graduate School of Life Sci., Univ. of Hyogo, [3]Graduate School of Frontier Sciences, U. Tokyo

The tunicate *Ciona intestinalis*, an invertebrate chordate, has recently emerged as a powerful model organism for gene regulation analysis. However, few studies have been conducted to identify and characterize its transcription start sites (TSSs) and promoters at the genome-wide level. Here, using TSS-seq, we identified TSSs at the genome-wide scale and characterized promoters in *C. intestinalis*. Specifically, we identified TSS clusters (TSCs), high-density regions of TSS-seq tags, each of which appears to originate from an identical promoter. TSCs were found not only at known TSSs but also in other regions, suggesting the existence of many unknown transcription units in the genome. We also identified candidate promoters of 79 ribosomal protein (RP) genes, each of which had the major TSS in a polypyrimidine tract and showed a sharp TSS distribution like human RP gene promoters. *Ciona* RP gene promoters, however, did not appear to have typical TATA boxes unlike human RP gene promoters. Surprisingly, despite the absence of CpG islands, *Ciona* TATA-less promoters showed low expression specificity like CpG-associated human TATA-less promoters. Using TSS-seq, we also predicted non-operon-type *trans*-spliced gene TSSs, and found that their downstream regions had higher G + T content than those of non-*trans*-spliced gene TSSs. This higher G + T content was also observed in downstream regions of operon-type *trans*-spliced gene TSSs. Although the mechanism of *trans*-splicing remains unclear, the conservation of high G + T content in two different types of *trans*-spliced genes may suggest its importance in *C. intestinalis trans*-splicing. Our results provide valuable information about TSSs and promoter characteristics in *C. intestinalis* and will be helpful in future analysis of transcriptional regulation in chordates.

2. OpenTein: a database of digital whole-slide images of stem cell-derived teratomas

Sung-Joon Park, Yusuke Komiyama, Hirofumi Suemori[4], Akihiro Umezawa[5] and Kenta Nakai:

[4]Department of Embryonic Stem Cell Research, Institute for Frontier Medical Sciences, Kyoto University, [5]Department of Reproductive Biology, National Institute for Child Health and Development

Human stem cells are promising sources for regenerative therapy. To ensure safety of future therapeutic applications, the differentiation potency of stem cells has to be tested and be widely opened to the public. The potency is generally assessed by teratoma formation comprising differentiated cells from all three germ layers, and the teratomas can be inspected through high-quality digital images. The teratoma assay, however, lacks consistency in transplantation protocols and even in interpretation, which needs community-based efforts for improving the assay quality. Here, we have developed a novel database OpenTein (Open Teratoma Investigation, http://opentein.hgc.jp/) to archive and freely distribute high-resolution whole-slide images and relevant records. OpenTein has been designed as a searchable, zoomable and annotatable web-based repository system, and users can freely access and process the digital teratoma images. Our system offers valuable tools and resources in the new era of stem cell biology.

3. Transcriptome characterization of diverse stem cells cultured at various laboratories

Sung-Joon Park and Kenta Nakai

Human stem cells are becoming more and more essential for cell biology and regenerative therapy. One of the major issues facing the communities is to establish standard protocols for stem cell culture, which confers the homogeneous potency of stem cell lines on multiple facilities. To address this issue, we performed whole-transcriptome analysis with more than 50 stem cell samples collected from multiple laboratories where different culturing procedures for a cell line were used. The stem cells included human embryonic stem cells (hESCs), induced pluripotent stem cells (hiPSCs) and mesenchymal stem cells (hMSCs). The stem cells were cultured up to ～140 passages under various conditions; with or without feeder cells (MEF or SNL), hypoxia (5%) or normal (20%) oxygen density, Primate ES medium or DMEM/F-12. This study revealed that 7-25% of total sequenced reads in samples with feeder cells were contaminants (mouse RNAs), which was higher than that in feeder-free samples (0.18%). Importantly, our transcriptome profile clearly grouped sample providers rather than cell lines, highlighting the importance of quality-control procedure for cultured stem cell lines. Particularly, RNA-seq data with samples cultured on normal oxygen density captured ectopic gene expression related to cell differentiation, mitochondrial activity, and oxidative stress response. This may reflect that stem cells reside in niches characterized by hypoxia and low reactive oxygen species. In addition, the profile captured genes differentially expressed in stem cells; e.g. the expression of CD106 was specific to hMSCs while pluripotent marker genes (NANOG, SOX2 and POU5F1) were observed only in pluripotent stem cells. Our transcriptome characterization offers valuable resources in standardizing stem cell culture.

4. Development of Open Source Web application platforms of image annotation viewer

Yusuke Komiyama, Sung-Joon Park, Mihoko Saito-Adachi[6], Emi Ikeda and Kenta Nakai: [6]Division of Cancer Genomics, National Cancer Center Research Institute.

Various types of image data are generated and accumulated in biomedical science every day. We have developed and released a lightweight and portable Web-based IAV platform for biomedical images, which can store and share annotations as metadata while maintaining the original image. The OpenIAV platform can be used to annotate and share experimental results and can facilitate the development of a laboratory note archive to ensure that legal regulatory requirements are met. Also, the IAV platform can be used to create publicly available biological image databases. Besides, we designed the resource description framework (RDF) scheme for setting the metadata of annotation of an IAV platform to RDF using extensible markup language (XML) in a temporary way.

5. Development of Open Source Web application platform and standard operating procedure for Laboratory Information Management System

Yusuke Komiyama, Sung-Joon Park, Mihoko Saito-Adachi[6], Emi Ikeda and Kenta Nakai

Scientists and doctors need Standard Operating Procedure (SOP) managerial software for planning experimental methods. SOP Open source software that satisfies these requirements does not exist yet. We propose a novel Web application platform of SOP that is structured by SOP server and SOP client for a tablet PC such as iPad. OpenSOP can quickly archive a SOP and share it on the Web browser. It can easily install on a user's PC or server machine from the Docker container to build a Web application in the cloud computing environment. It can run on the 64bit OS of Windows, Mac OS X and Linux. The source code and Docker containers of the OpenIAV and OpenSOP platforms have been made available (http://regmed.hgc.jp/).

6. An Integrative Approach for Efficient Analysis of Whole Genome Bisulfite Sequencing Data

Jong-Hun Lee, Sung-Joon Park and Kenta Nakai

Whole genome bisulfite sequencing (WGBS) is a high-throughput technique for profiling genome-wide DNA methylation at single nucleotide resolution. However, the applications of WGBS are limited by low accuracy resulting from the falsely mapped reads. We tried to improve the accuracy by analyzing and integrating the performances of the three most widely used bisulfite-read mappers: Bismark, BSMAP and BS-seeker2. A comprehensive analysis of the three mappers revealed that the mapping results of the mappers were mutually complementary under diverse read conditions. Therefore, we sought to integrate the characteristics of the mappers by scoring them to gain robustness against artifacts. As a result, the integration significantly increased detection accuracy compared with the individual mappers. In addition, the amount of detected cytosine was higher than that by Bismark. Furthermore, the integration successfully reduced the fluctuation of detection accuracy induced by read conditions. We applied the integration to real WGBS samples and succeeded in classifying the samples according to the originated tissues by both CpG and CpH methylation patterns. This study contributes to DNA methylation research by improving efficiency of methylation detection from WGBS data and facilitating the comprehensive analysis of public WGBS data.

7. A study on the application of Topic Models to motif finding

Josep Basha Gutierrez and Kenta Nakai

The discovery of sequence elements that are bound by DNA-binding proteins is a fundamental problem in understanding transcriptional regulation. To facilitate the task of identifying these elements, called transcription factor binding sites (TFBS), several computational approaches have been applied. In this study, we analyze the application of a novel strategy based on the use of Topic Models. These are statistical models which parse the distribution of the words of a given vocabulary in a collection of documents in order to identify the abstract "topics" that occur in them. We applied this model to a collection of sequences with the goal of finding motifs. The results, though did not especially stand out in contrast with other methods, suggest that the approach is accurate enough to be considered a valid motif finding tool. After combining the new method with our previous algorithm based on the use of a genetic algorithm and statistical coefficients, the results clearly outperformed all of the other methods studied in sensitivity and in overall performance at site level, showing also acceptable rates in all of the other statistics. This further proves that the mixed algorithm can be a powerful tool to successfully predict motifs in different kinds of sets of DNA sequences.

8. Update of databases Hintdb and HitPredict

Yosvany López, Kenta Nakai and Ashwini Patil

Protein-protein interactions (PPIs) are vital for cellular function in organisms and hence their detection is of considerable importance. The advent of high-throughput technologies has led to a manifold increase in the PPI information. HitPredict (http://hintdb.hgc.jp/htp) is a consolidated resource of experimentally identified, physical protein-protein interactions with confidence scores to indicate their reliability. The study of genes and their inter-relationships using methods such as network and pathway analysis requires high quality protein-protein interaction information. Extracting reliable interactions from most of the existing databases is challenging because they either contain only a subset of the available interactions, or a mixture of physical, genetic and predicted interactions. Automated integration of interactions is further complicated by varying levels of accuracy of database content and lack of adherence to standard formats. To address these issues, the latest version of HitPredict provides a manually curated dataset of 398,696 physical associations between 70,808 proteins from 105 species. Manual confirmation was used to resolve all issues encountered during data integration. For improved reliability assessment, this version combines a new score derived from the experimental information of the interactions with the original score based on the features of the interacting proteins. The combined interaction score performs better than either of the individual scores in HitPredict as well as the reliability score of another similar database. HitPredict provides a web interface to search proteins and visualize their interactions, and the data can be downloaded for offline analysis. Data usability has been enhanced by mapping protein identifiers across multiple reference databases. Thus, the latest version of HitPredict provides a significantly larger, more reliable and usable dataset of protein-protein interactions from several species for the study of gene groups. Similarly, Hintdb (http://hintdb.hgc.jp/hint/), a database of homologous protein-protein interactions was also updated.

9. Identification of pathogen-specific response pathways in activated immune cells using a systems biology approach

Ashwini Patil and Kenta Nakai

The innate immune response is the first level of protection in organisms against invading pathogens. It is primarily mediated by the Toll-like receptors functioning through the Myd88-dependent and TRIF-dependent pathways. Despite being widely studied, it is not yet completely understood and systems-level analyses have been lacking. In this study, we identified high-probability networks of genes activated during the innate immune response on exposure to five pathogenic components. We used TimeXNet, a tool using a network flow optimization approach, to analyze time course gene expression profiles of activated immune cells in the context of a large gene regulatory and protein-protein interaction network. We compared the regulatory networks responsible for the distinct immune outcomes produced from different pathogens to identify unique regulatory genes associated with each pathogenic component.

10. Modeling the cis-regulatory modules of genes expressed in the developmental cycle of *Drosophila melanogaster*

Yosvany López, Alexis Vandenbon[7] and Kenta Nakai: [7]WPI-iFReC, Osaka University

The understanding of transcription as the first step in the biological regulation process has become absolutely necessary for comprehensively studying the expression of genes. During transcription, DNA-binding proteins modulate the expression of genes by binding to specific regulatory elements in nearby genomic regions. It has been hypothesized that the promoters of genes expressed in the same tissue or cell type might somehow share common structural patterns. We have conducted an extensive study on the cis-regulatory modules of genes expressed in several developmental stages of Drosophila melanogaster. A collection of six types of structural features related to positioning and presence of motifs, pairwise distance between motifs, positioning of motif pairs to the transcription start site, order and orientation was proposed. Stage-related motifs predicted for each set of expressed genes were used to compute the previous features. Those highly informative characteristics were subsequently used for searching stage-expressed genes with similar promoter architecture in the entire set of *D. melanogaster* regulatory regions. RNA-sequencing data from the modENCODE project were finally utilized for validating the different models.

11. Clustering analysis of alternative splice sites using RNA-seq samples of multiple tissues and species

Shunichi Wakabayashi and Kenta Nakai

Elucidation of the mechanism of alternative RNA splicing in eukaryotic cells is important but still unclear. The categorization of splice sites which are regulated by common splicing mechanisms can help in understanding this process. We developed a novel method to make clusters of splice sites using RNA-seq data of multiple tissues and samples. In the method, we defined feature vectors and distances between two vectors which represent variation of splice patterns of observed alternative splice sites across multiple samples and their similarities. The distances of the feature vectors can be used as input of non-hierarchical clustering method and the result of clustering consists of groups of alternative splice sites which have a similar splicing pattern. We validated the method and analyzed its result using RNA-seq data of various kinds of somatic tissues of human and mouse released from ENCODE project. The result of clustering showed that each cluster contained the similar tissue specific alternative splice sites and these sites shared some common biological properties. For example, according to motif searching, RBFOX1 binding motif was enriched nearby nervous cells and heart specific splice sites in agreement with previous studies. Other analyses about gene functions and exon-intron structures of each cluster suggested new functions and properties of alternative splicing. We could confirm that our method is useful to reveal new knowledge about alternative splicing from large scale transcriptome data.

12. Analysis of genome stability in patient iPSC generation and TALEN genome editing

Sung-Joon Park, Takefumi Sone[8], Kuo-ching Liang, Hideyuki Okano[8] and Kenta Nakai: [8]Dept. of Physiology, Graduate School of Medicine, Keio University

In cell replacement therapy using pluripotent stem cells, establishing cell culture and differentiation methods that reduce the risk of viral or oncogenic activity is key for its success. This still needs community effort. In this research activity, we intend to address the possibility of high-throughput genome sequencing as an evaluation tool of genome stability in iPSC generation and its disease modeling. Here, we applied genome sequencing technologies to two patient-derived iPS cell lines; whole-genome sequencing (WGS) for Dravet syndrome patient-derived iPSCs before and after TALEN genome editing and whole-exome sequencing (WES) for Parkinson's disease patient-derived iPSCs before and after TALEN genome editing. We first confirmed that the disease-causing mutations were successfully corrected by TALEN system; the

mutation of C to T in exon 26 of SCN1A (Dravet syndrome) and the mutation of C to T in exon 41 of LRRK2 (Parkinson's disease) were repaired, respectively. Next, we developed a pipeline of paired-end read mapping to identify the genomic positions where non-human DNA fragments are inserted. Using the pipeline, we confirmed that the viral cassette, a component of TALEN donor vector that is deliberately retained in the iPSCs, exists at the target-site only, which suggests the high site-specificity of TALEN system without off-target effects. In addition, the pipeline detected the integration sites of viral vectors that were used in iPSC generation, and revealed that the sites are present within intergenic positions. Remarkably, in the case of Dravet syndrome patient-derived iPSCs, single nucleotide variations of high confidence were found from over 400 loci in protein-coding exons. Among them, 22 loci were uniquely found from iPSCs after genome editing. Further analysis of functional characterization of those variations will give a clue to efficiently maintain genome stability during iPSC generation, which promotes the clinical application of regenerative medicine.

Publications

Park, S.J., Shirahige, K., Ohsugi, M., Nakai, K. DBTMEE: a database of transcriptome in mouse early embryos. *Nucl. Acids Res.* 43: D771-D776, 2015.

Suzuki, A., Wakaguri, H., Yamashita, R., Kawano, S., Tsuchihara, K., Sugano, S., Suzuki, Y., Nakai, K. DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data. *Nucl. Acids Res.*, 43: D87-D91, 2015.

Gutierrez, J.B., Frith, M., Nakai, K. A genetic algorithm for motif finding based on statistical significance. *Bioinformatics and Biomedical Engineering* (Third International Conference, IWBBIO 2015, Granada, Spain, April 15-17, 2015, Proceedings, Part I), Springer, *LNCS* 9043: 438-449, 2015.

Lee, J.H., Park, S.J., Nakai, K. An integrative approach for efficient analysis of whole genome bisulfite sequencing data. *BMC Genomics* 16(Suppl 12): S14, 2015.

Liang, K.C., Patil, A., Nakai, K. Discovery of intermediary genes between pathways using sparse regression. *PLoS One*, 10(9): e0137222, 2015.

López, Y., Nakai, K., and Patil, A. HitPredict version 4 - comprehensive reliability scoring of physical protein-protein interactions from more than 100 species. DATABASE 2015: bav117, 2015.

Srihari, S., Yong, C.H., Patil, A., Wong, L. Methods for protein complex prediction and their contributions towards understanding the organization, function and dynamics of complexes. FEBS Letters, 589(19 Pt A): 2590-602, 2015.

Park, S.J., Komiyama, Y., Suemori, H., Umezawa, A., Nakai, K. OpenTein: a database of digital whole-slide images of stem cell-derived teratomas. *Nucl. Acids Res.*, 44(D1): D1000-D1004, 2016.

Yokomori, R., Shimai, K., Nishitsuji, K., Suzuki, Y., Kusakabe, K.G., Nakai, K. Genome-wide identification and characterization of transcription start sites and promoters in the tunicate *Ciona intestinalis. Genome Res.*, 26: 140-150, 2016.

Park, S.J., Saito-Adachi, M., Komiyama, Y., Nakai, K. Advances, practice, and clinical perspectives in high-throughput sequencing, *Oral Diseases*, in press.

Sasamoto, Y., Hayashi, R., Park, S.-J., Saito-Adachi, M., Suzuki, Y., Kawasaki, S., Quantock, A., Nakai, K., Tsujikawa, M., and Nishida, K., PAX6 isoforms, along with reprogramming factors, differentially regulate the induction of cornea-specific genes. *Sci. Rep.*, 6: 20807, 2016.

Human Genome Center

Department of Public Policy
公共政策研究分野

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

The Department of Public Policy contributes to achieve three major missions: public policy science studies of translational research and its impact on society; research ethics consultation for scientists to comply with ethical guidelines and to build public trust; and development of "minority-centered" scientific communication. By conducting qualitative and quantitative social science study and policy analysis, we facilitate discussion of challenges arising from advances in medical sciences.

1. Research ethics consultation and studies on ethical, legal, and social implications on genomic medicine

Japan Agency for Medical Research and Development (AMED) has commissioned to provide research ethics consultation to several large projects promoting genomic medicine, including The Biobank Japan (BBJ) Project (BBJP) and Project for Development of Innovative Research on Cancer Therapeutics (P-DIRECT). BBJP is a disease-focused biobanking project started in 2003. BBJ consists of donated DNA, sera, and clinical information from 200,000 patients in Japan, and continues to collect new DNA and clinical information. Informed consent, which ensures the autonomous decisions of participants, is believed to be practically impossible for the biobanking project in general. We issued semiannual newsletters for sample donors for transparency and information. Since 2014, BBJ started to store samples from National Hospital Organization (NHO), Japan Clinical Oncology Group (JCOG), and Japan Children's Cancer Group (JCCG). We supported to establish ethical policies for collecting

Figure 1. Examples of BBJP brochures we developed for sample donors and their physicians

these samples.

On the other hands, P-DIRECT promotes strategic research and development (R & D) of the basic compounds (seeds) that contribute to development of next-generation innovative diagnostic techniques and new therapeutic agents incorporating basic research results. We provided research ethics consultation for 34 research projects at 64 designated institutions. We advised those principal investigators to comply with ethical guidelines and provided sample consent forms. We held a forum for research ethics consultants in December of 2015 to review our consultation work.

2. Research ethics consultation and studies on ethical, legal, and social implications of stem cell research

Japan Agency for Medical Research and Development (AMED) has also commissioned us to provide research ethics consultation to stem cell research since 2012. The program is called "research on the ethical, legal, and social implications related to regenerative medicine". In order to make regenerative medicine more concrete, it is essential to promote research development with a definite focus on clinical applications and to establish a framework for clinical research at an early stage. In 2015, we provided more than 50 consultations for stem cell researchers. Topics of those consultations include research design, informed consent, IRBs, return of research results, inclusion criterion of participants of first-in-human trials and governance of iPSC banking.

Based on our training workshops informed consent (IC) for members of stem-cell research teams and an educational workshop for a certain patients' group, we published a paper to discuss necessary aspects in promoting and conducting IC appropriately. Additionally, it emphasizes the importance of continuous education for improving the IC skills and competency of people responsible for obtaining IC.

We also organized interdisciplinary research groups to address the ethical, legal, and social implications (ELSI) related to regenerative medicine in

Figure 2. Examples of brochures and informed consent tools of the program for intractable diseases research using disease-specific induced pluripotent stem (iPS) cells

a comprehensive manner, with a view to establishing a framework for ethical support and review of regenerative medicine.

3. Informed consent in clinical trials using stem cells: Suggestions and points of attention from informed consent training workshops in Japan

Informed consent (IC) is an essential requirement of ethical research involving human participants, and usually is achieved by providing prospective research participants (PRPs) with a document that explains the study and its procedures. However, results of a series of IC workshops held in Tokyo during 2014 indicate that consent forms alone are not enough to achieve full IC in regenerative medicine research due to the necessity of long-term patient-safety observations to meet the ethical challenges of such research. Adequate training of the people who are responsible for obtaining IC (elucidators) is also necessary to ensure full IC. Elucidators must be able to provide PRPs with sufficient information to assure adequate comprehension of the study and its potential aftereffects; judge PRPs' voluntariness and eligibility; and establish/maintain partnerships with PRPs. The workshops used role-playing simulations to demonstrate how to effectively obtain fuller IC to members of several Japanese research groups preparing for clinical stem cell trials. Workshop results were correlated with the results of a 2013 workshop on what information is patients want when considering participation in iPSC research. The correlated results showed the need for continuous training and education of elucidators in order to have them acquire and maintain IC competency.

4. Noncompliance with Human Subjects' Protection Requirements as a Reason for Retracting Papers

Though protection of human research subjects is universally recognized as a critical requirement for the ethical conduct of research, few studies have examined retractions of medical articles through apparent noncompliance with that requirement. From our survey of 99 retracted papers published from 1981 to 2011, we found that the basis for those decisions was poorly explained in retraction notices and that most of the articles continued to be cited. In retraction notices, the current manner of explaining failure to protect human subjects is misleading and confusing.

5. Public attitude towards governance of consumer targeted genetic testing and secondary use of research in Japan

In 2014, several Japanese industries announced the launch of DTC (direct-to-consumer) personal genetic testing services in Japan. The Ministry of Trade and Industry (METI) has encouraged an industrial association, Council for Protection of Individual Genetic Information (CPIGI), to prepare an accreditation system based on their best practice guidelines. On the other hands, the Ministry of Health, Labour and Welfare started to discuss the definitions of "diagnosis" to regulate DTC personal genetic testing services. Currently, the Japanese Diet is revising the Act on Protection of Personal Information in which personal genome data hasn't been covered. However, we still do not have enough data on public attitudes towards these services and regulation.

We conducted a cross-sectional and anonymous online surveys were administrated to men and women in Japan in March 2014 and March 2015. Participants of these studies have been extracted so as not to overlap from the survey panel. The questionnaires included questions concerning genetic knowledge, attitudes towards genetic testing and regulations governing DTC personal genetic testing services. We compared the data of these surveys and past relevant dataset.

The mean age of the respondents was 45.7 ± 14.0 years in 2014 and 45.8 ± 14.0 years in 2015. The percentage of the respondents who haven't known the term "genome" was lower than "gene", "DNA" and "iPS Cell". The percentage of the respondents who knew about DTC personal genetic testing services was 32.9% in 2015, which is higher than in the 2014 survey by 10 points. The percentage who were willing to purchase or who had purchased these services was almost the same level. The results of willingness to undergo 6 types of genetic testing by scientific evidence. More respondents in 2015 (35.4%) allowed the secondary use of genomic data for medical research than in 2014 (28.6%). The percentage of respondents who wanted to receive personal results were 34.3% in 2014 and 33.6% in 2015. The Japanese public showed slightly more positive attitudes towards DTC personal genetic testing services and the secondary use for medical research. However, more than 50% of our respondents couldn't understand the meaning of scientific evidences. There may be a possibility of affecting the attitude about the future of the services.

6. Patients' perspective and experiences of clinical trials in Japan: return of research results

While patients' involvement is essential to improve systems and circumstances of clinical trials, their experiences rarely be shared with others. Additionally, few national surveys on patients' attitudes toward clinical trials and their experiences

have been conducted. The objective of this study is to investigate patients' perspectives and experiences of clinical trials in Japan.

We conducted a cross-sectional and web-based questionnaire survey in March 2014. We also referred our interview dataset from 35 patients which were by applying the same methods used by the Oxford Health Experiences Research Group.

We analyzed a dataset obtained from 12,506 patients (Response rate = 58%). Mean age of the respondents was 52 (range, 20 to 79). "Randomization" and "placebo" were poorly understood (25.5% and 56.1%, respectively) while most of them (87.4 %) understood that clinical trials were conducted to establish evidence for future medicine. Of the respondents, 8.7% (n = 967) had enrolled a clinical trial. The major motivations for enrolling were "contributing to develop new therapies for my disease" (44.2%) and "contributing to the advances in medicine" (42.7%). Among the enrollee, most of them (76.5%) have not known about the results of clinical trials.

The Declaration of Helsinki state that "Negative and inconclusive as well as positive results must be published or otherwise made publicly available". However, it is not strange their indifferent attitudes toward research results if the enrollee hoped for health benefit due to therapeutic misconception (Appelbaum et al. 1982) or a rational wager (Locock and Smith 2011).

Publications

1. Kusunose M, Nagamura F, Muto K. Informed Consent in Clinical Trials Using Stem Cells: Suggestions and Points of Attention from Informed Consent Training Workshops in Japan. S Afr J BL 2015; 8 (2 Suppl 1): 49-54. DOI: 7196/SAJBL.8016
2. Inoue Y, Muto K. Noncompliance with Human Subjects' Protection Requirements as a Reason for Retracting Papers: Survey of Retraction Notices on Medical Papers Published from 1981 to 2011. Account Res. 2016; 23(2): 123-35.
3. 神里彩子・武藤香織編『医学・生命科学の研究倫理ハンドブック』，東京大学出版会，2015.
4. 井上悠輔(分担)『公衆衛生倫理』(「第３章　公衆衛生の歴史：日本(1)概論」「第４章　公衆衛生の歴史：日本(2)主な関連法規」，赤林朗・児玉聡編)，勁草書房，2015.