Health Intelligence Center

Division of Health Medical Data Science
健康医療データサイエンス分野

Professor	Seiya Imoto, Ph.D.	教　授　博士(数理学)	井　元　清　哉
Assistant Professor	Takanori Hasegawa, Ph.D.	助　教　博士(情報学)	長谷川　嵩　矩

Our mission is to utilize genomic big data and time series health medical data to realize methods for prediction and prevention of diseases and keeping/improving our health. For this purpose, we develop novel computational data analysis technologies by integrating Bayesian statistical theory and high-performance computing on supercomputer system.

1. Development of Computational Platform for Clinical Sequence and Interpretation

Shimizu E, Yui H, Kasajima R, Yamaguchi K, Yokoyama K, Komura M, Saito A, Kobayashi M, Yuji K, Shibuya T, Hasegawa T, Niida A, Miyagi Y, Muto K, Tojo A, Furukawa Y, Imoto S, Miyano S, Yamaguchi R

From April 2015, Medical Genomics Research Initiative The University of Tokyo is launched. For implementing clinical sequence in the Institute of Medical Science, we formed a team of researchers and technicians who have various academic backgrounds including medicine, biology, pharmacology, genetics, statistics, computer science, ethics, etc. A highly secure infrastructure for analyzing personal genome was constructed; in the space, next generation sequencers are directly connected to a part (disconnect to internet) of supercomputer system in Human Genome Center and, for keeping traceability, laboratory information management system (LIMS) is installed to record all logs of wet experiments and computational analyses. Together with genome analysis in clinical sequence, we now intensively focus on a method for interpreting personal genome information. In July 2015, we started to use IBM Watson for cancer research to interpret the results of genome analyses. The results of genome sequence analysis including the interpretation of IBM Watson are evaluated and discussed in biweekly sequence board meeting. In 2016, we analyzed around sequence data of 100 cancer patients (more than 250 sequencing samples) with whole genome, exome, target deep sequencings. Also, multi-omics data including genome, transcriptome and epigenome were measured for integrative analysis that has the potential to achieve highly precise interpretation. This research is also performed as a part of the University of Tokyo's Center of Innovation (COI) project "Self-Managing Healthy Society".

2. Health Medical big data analysis

a. Integration of the records of health examination, microbiome and genomic data for predicting disease risks

Hasegawa T, Kakuta M, Niida A, Yamaguchi R, Imoto S

To predict disease risk, we investigate a computational method to integrate the data of health examination, microbiome and genome. We collaborate with the research project at Hirosaki University

COI program "The Center of Healthy Aging Innovation" who collects those data of more than 20 thousand participants in total over 10 years. Our IMSUT team analyzes their data on the supercomputer system Shirokane3.

3. Computational Methods in Systems Biology and Immunology

a. Interaction-Based Feature Selection for Cancer Driver Genes Detection Through Copy Number-Driven Expression Level.

Park H[1], Niida A, Imoto S, Miyano S: [1]Faculty of Global and Science Studies, Yamaguchi University

Driver gene selection is crucial to understand the heterogeneous system of cancer. To identity cancer driver genes, various statistical strategies have been proposed, especially the L1-type regularization methods have drawn a large amount of attention. However, the statistical approaches have been developed purely from algorithmic and statistical point, and the existing studies have applied the statistical approaches to genomic data analysis without consideration of biological knowledge. We consider a statistical strategy incorporating biological knowledge to identify cancer driver gene. The alterations of copy number have been considered to driver cancer pathogenesis processes, and the region of strong interaction of copy number alterations and expression levels was known as a tumor-related symptom. We incorporate the influence of copy number alterations on expression levels to cancer driver gene-selection processes. To quantify the dependence of copy number alterations on expression levels, we consider [Formula: see text] and [Formula: see text] effects of copy number alterations on expression levels of genes, and incorporate the symptom of tumor pathogenesis to gene-selection procedures. We then proposed an interaction-based feature-selection strategy based on the adaptive L1-type regularization and random lasso procedures. The proposed method imposes a large amount of penalty on genes corresponding to a low dependency of the two features, thus the coefficients of the genes are estimated to be small or exactly 0. It implies that the proposed method can provide biologically relevant results in cancer driver gene selection. Monte Carlo simulations and analysis of the Cancer Genome Atlas (TCGA) data show that the proposed strategy is effective for high-dimensional genomic data analysis. Furthermore, the proposed method provides reliable and biologically relevant results for cancer driver gene selection in TCGA data analysis.

b. Sequence-specific bias correction for RNA-seq data using recurrent neural networks

Zhang Y-Z, Yamaguchi R, Imoto S, Miyano S.

The recent success of deep learning techniques in machine learning and artificial intelligence has stimulated a great deal of interest among bioinformaticians, who now wish to bring the power of deep learning to bare on a host of bioinformatical problems. Deep learning is ideally suited for biological problems that require automatic or hierarchical feature representation for biological data when prior knowledge is limited. In this work, we address the sequence-specific bias correction problem for RNA-seq data redusing Recurrent Neural Networks (RNNs) to model nucleotide sequences without pre-determining sequence structures. The sequence-specific bias of a read is then calculated based on the sequence probabilities estimated by RNNs, and used in the estimation of gene abundance. We explore the application of two popular RNN recurrent units for this task and demonstrate that RNN-based approaches provide a flexible way to model nucleotide sequences without knowledge of predetermined sequence structures. Our experiments show that training a RNN-based nucleotide sequence model is efficient and RNN-based bias correction methods compare well with the-state-of-the-art sequence-specific bias correction method on the commonly used MAQC-III data set. RNNs provides an alternative and flexible way to calculate sequence-specific bias without explicitly pre-determining sequence structures.

c. Bayesian model for analyzing human leukocyte antigen regions

Hayashi S, Yamaguchi R, Mizuno S[3], Komura M, Miyano S, Nakagawa H[4], Imoto S: [3]Center for Advanced Medical Innovation, Kyushu University, [4]RIKEN Center for Integrative Medical Sciences

Although human leukocyte antigen (HLA) genotyping based on amplicon, whole exome, and RNA sequence data has been achieved in recent years, accurate genotyping from whole genome sequence (WGS) data remains a challenge due to the low depth. Furthermore, there is no versatile method for in- depth analysis of HLA genes. We developed ALPHLARD, which is a Bayesian model that collects reads potentially generated from HLA genes and accurately determines a pair of HLA types for each of HLA-A, -B, -C, -DPA1, - DPB1, -DQA1, -DQB1, and -DRB1 genes at 6-digit resolution. Furthermore, ALPHLARD can detect rare germline variants not stored in HLA databases, reconstruct unknown non-coding sequences, and call somatic mutations from paired normal and tumor WGS data. We illustrate the capability of ALPHLARD using 25 WGS data sets from an Illumina platform. By comparing the results of HLA genotyping from SBT and am-

plicon sequencing methods, ALPHLARD achieved 98.1% accuracy at 6-digit resolution. Moreover, we could reconstruct 99.92% of unknown non-coding sequences of HLA-A*26:03:01. We also detected three somatic point mutations and one case of loss of heterozygosity in the HLA genes of the samples. This research is performed as a part of International Cancer Genome Consortium PanCancer Analysis Project.

d. An *in silico* automated pipeline to identify tumor specific neoantigens from next generation sequencing data

Hasegawa T, Hayashi S, Shimizu E, Mizuno S, Yamaguchi R, Miyano S, Nakagawa S, Imoto S:

Recent progress of massive parallel sequencing technology enables us to detect somatic mutations in each of cancer patients. It is known that some mutated peptides produced from missense mutations binds to the major histocompatibility complex (MHC). Since MHC presents mutated peptides to anti-tumor T cells, understanding this process is important in cancer immunotherapy. In this paper, we introduce a computational pipeline to predict binding affinity between mutated peptides and MHC molecules to detect neoantigens. We have implemented this pipeline on our supercomputer system. With nonsynonymous substitutions, frameshift insertions and deletions detected and intron retentions from whole-genome or exome sequencing data, we utilize RNA sequencing data and annotation data to make neoantigen detection pipeline more accurate.

e. Somatic mutation detection from next generation sequencing data

Moriyama T, Shiraishi Y, Chiba K, Yamaguchi R, Imoto S, Miyano S

Detection of somatic mutations from tumor and matched normal sequencing data has become a standard approach in cancer research. Although a number of mutation callers are developed, it is still difficult to detect mutations with low allele frequency even in exome sequencing. We expect that overlapping paired-end read information is effective for this purpose, but no mutation caller has modeled overlapping information statistically in a proper form in exome sequence data. Here, we develop a Bayesian hierarchical method, OVarCall (https://github.com/takumorizo/OVarCall), where overlapping paired-end read information improves the accuracy of low allele frequency mutation detection. Firstly, we construct two generative models: one is for reads with somatic variants generated from tumor cells and the other is for reads that does not have somatic variants but potentially includes sequence errors. Secondly, we calculate marginal likelihood for each model using a variational Bayesian algorithm to compute Bayes factor for the detection of somatic mutations. We empirically evaluated the performance of OVarCall and confirmed its better performance than other existing methods.

Publications

1. VanderWeele DJ, Finney R, Katayama K, Gillard M, Paner G, Imoto S, Yamaguchi R, Wheeler D, Cam M, Maejima K, Sasaki-Oku A, Nakano K, Tanaka H, Pontier A, Grigoryev D, Kubo M, Ratain M, Miyano S, Nakagawa H. Local enrichment of multiple clinically significant genetic alterations within index foci of potentially lethal prostate cancer, *European Urology Focus,* in press.
2. Fujita K, Chen X, Homma H, Tagawa K, Amano M, Saito A, Imoto S, Akatsu H, Hashizume Y, Kaibuchi K, Miyano S, Okazawa H. Targeting Tyro3 signals ameliorates PGRN-mutant FTLD-TDP model via tau-mediated synapse pathology, *Nature Communications,* 9: 433, (2018)
3. Ogawa M, Yokoyama K, Hirano M, Jimbo K, Ochi K, Kawamata T, Ohno N, Shimizu E, Yokoyama N, Yamaguchi R, Imoto S, Uchimaru K, Takahashi N, Miyano S, Imai Y, Tojo A. Different clonal dynamics of chronic myeloid leukaemia between bone marrow and the central nervous system, *British Journal of Haematology,* in press.
4. Zhang Y-Z, Imoto S, Miyano S, Yamaguchi R. Reconstruction of high read-depth signals from low-depth whole genome sequencing data using deep learning, *International Workshop on Deep Learning in Bioinformatics, Biomedicine, and Healthcare Informatics,* 1227-1232 (2017) doi: 10.1109/BIBM.2017.8217832
5. Kobayashi M, Yokoyama K, Shimizu E, Yusa N, Ito M, Yamaguchi R, Imoto S, Miyano S, Tojo A. Phenotype-based gene analysis allowed successful diagnosis of X-linked neutropenia associated with a novel WASp mutation. *Ann Hematol.,* (2017) Sep 27. doi: 10.1007/s00277-017-3134-3. [Epub ahead of print]
6. Kiyotani K, Mai T, Yamaguchi R, Yew P-Y, Kulis M, Orgel K, Imoto S, Miyano S, Burks A. W, Nakamura Y. Characterization of the B-cell receptor repertoires in peanut allergic subjects

undergoing oral immunotherapy, *Journal of Human Genetics,* in press.
7. Sato R, Shibata T, Tanaka Y, Kato C, Yamaguchi K, Furukawa Y, Shimizu E, Yamaguchi R, Imoto S, Miyano S, Miyake K. Requirement of glycosylation machinery in TLR responses revealed by CRISPR/Cas9 screening, *International Immunology,* in press.
8. Tanikawa C, Zhang YZ, Yamamoto R, Tsuda Y, Tanaka M, Funauchi Y, Mori J, Imoto S, Yamaguchi R, Nakamura Y, Miyano S, Nakagawa H, Matsuda K. The Transcriptional Landscape of p 53 Signalling Pathway. *EBioMedicine,* 2017 Jun; 20: 109-119.
9. Miyamoto T, Tanikawa C, Yodsurang V, Zhang Y-Z, Imoto S, Yamaguchi R, Miyano S, Nakagawa H, Matsuda K. Identification of a p53-repressed gene module in breast cancer cells, *Oncotarget,* 2017 Jul 26; 8(34): 55821-55836.
10. Nagata H, Kozaki K, Muramatsu T, Hiramoto H, Tanimoto K, Fujiwara N, Imoto S, Ichikawa D, Otsuji E, Miyano S, Kawano T, Inazawa J. Genome-wide screening of DNA methylation associated with lymph node metastasis in esophageal squamous cell carcinoma, *Oncotarget,* 8(23): 37740-37750.
11. Fujii K, Miyahara Y, Harada N, Muraoka D, Komura M, Yamaguchi R, Yagita H, Nakamura J, Sugino S, Okumura S, Imoto S, Miyano S, Shiku H. Identification of an immunogenic neoepitope encoded by mouse sarcoma using CXCR3 ligand mRNAs as sensors, *OncoImmunology,* 6(5): e1306617 (2017)
12. Moriyama T, Shiraishi Y, Chiba K, Yamaguchi R, Imoto S, Miyano S. OVarCall: Bayesian mutation calling method utilizing overlapping paired-end reads, *IEEE Transactions on NanoBioscience,* 16(2): 1-7 (2017) 10.1109/TNB.2017.2670601
13. Park H, Shiraishi Y, Imoto S, Miyano S. A novel adaptive penalized logistic regression for uncovering biomarker associated with anti-cancer drug sensitivity, *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 14: 771-782 (2017)
14. Tsuda Y, Tanikawa C, Miyamoto T, Hirata M, Yodsurang V, Zhang YZ, Imoto S, Yamaguchi R, Miyano S, Takayanagi H, Kawano H, Nakagawa H, Tanaka S, Matsuda K. Identification of a p53 target, CD137L, that mediates growth suppression and immune response of osteosarcoma cells. *Sci Rep.,* 2017 Sep 6; 7(1): 10739. doi: 10.1038/s41598-017-11208-x.
15. Zhang Y-Z, Yamaguchi R, Imoto, S, Miyano S, Sequence-specific bias correction for RNA-seq data using recurrent neural networks, *BMC Genomics* 18(Suppl 1): 1044 (2017).
16. Park H, Niida A, Imoto S, Miyano S, Interaction based feature selection for uncovering cancer driver genes via copy number driven expression level, *Journal of Computational Biology,* 24: 138-152 (2017)

Health Intelligence Center

Division of Health Medical Computational Science
健康医療計算科学分野

Professor	Satoru Miyano, Ph.D.	教　授　理学博士	宮　野　　悟
Assistant Professor	Atsushi Niida, Ph.D.	助　教　博士(理学)	新井田　厚　司

The mission of this division is to develop computational science for transforming biomedical data to knowledge. By making full use of supercomputers, we are now focusing on annotation, translation and interpretation of genomic data including RNA sequences for supporting cancer research and clinical sequence.

1. Computational Science for Cancer Research

a. Understanding intratumor heterogeneity by combining genome analysis and mathematical modeling

Niida A, Nagayama S[2], Miyano S, Mimori K[1]: [1]Kyushu University Beppu Hospital, [2]Cancer Institute Hospital

Cancer is composed of multiple cell populations with different genomes. Each of the populations is called a clone (or subclone) and this phenomenon is called intratumor heterogeneity (ITH). ITH is observed in various types of cancers and presumed to be a major cause leading to therapeutic resistance. If a tumor harbors a major clone sensitive to a specific anti-cancer treatment, the tumor shrinks within a given period after the treatment. However, in most cases, a minor clone resistant to the chemotherapy exists in the tumor and predominantly regrows in spite of the intensive therapy. It is supposed that ITH can be generated by clonal branching during cancer evolution.

b. Japanese genome-wide association study identifies a significant colorectal cancer susceptibility locus at chromosome 10p14

Takahashi Y[1], Sugimachi K[1], Yamamoto K[3], Niida A, Shimamura T[4], Sato T[5], Watanabe M[6], Tanaka J[7], Kudo S[7], Sugihara K[8], Hase K[9], Kusunoki M[10], Yamada K[11], Shimada Y[12], Moriya Y[12], Suzuki Y[13], Miyano S, Mori M[14], Mimori K[1]: [3]Kurume University, [4]Nagoya University School of Medicine, [5]Medical Institute of Bioregulation, Kyushu University, [6]Kitazato University, [7]Showa University, [8]Tokyo Medical and Dental University, [9]National Defense University, [10]Mie University, [11]Takano Hospital, [12]National Cancer Center, [13]Graduate School of Frontier Sciences, University of Tokyo, [14]Osaka University School of Medicine

Genome-wide association studies are a powerful tool for searching for disease susceptibility loci. Several studies identifying single nucleotide polymorphisms (SNP) connected intimately to the onset of colorectal cancer (CRC) have been published, but there are few reports of genome-wide association studies in Japan. To identify genetic variants that modify the risk of CRC oncogenesis, especially in the Japanese population, we performed a multistage genome-wide association study using a large number of samples: 1846 CRC cases and 2675 controls. We identified 4 SNP (rs7912831, rs4749812, rs7898455 and rs10905453) in chromosome region 10p14 associated with CRC; however, there are no

coding or non-coding genes within this region of fairly extensive linkage disequilibrium (a 500-kb block) on 10p14. Our study revealed that the 10p14 locus is significantly correlated with susceptibility to CRC in the Japanese population, in accordance with the results of multiple studies in other races.

c. Personalized management of pancreatic ductal adenocarcinoma patients through computational modeling

Yamamoto KN[15], Yachida S[16], Nakamura A[17], Niida A, Oshima M[18], De S[18], Rosati LM[19], Herman JM[19], Iacobuzio-Donahue CA[20], Haeno H[15]: [15]Kyushu University, [16]National Cancer Center Research Institute, [17]Massachusetts General Hospital, [18]Kagawa University, [18]University of Colorado School of Medicine, [19]Johns Hopkins University School of Medicine, [20]Memorial Sloan-Kettering Cancer Center

Phenotypic diversity in pancreatic ductal adenocarcinoma (PDAC) results in a variety of treatment responses. Rapid autopsy studies have revealed a subgroup of PDAC patients with a lower propensity to develop metastatic disease, challenging the common perception that all patients die of widely metastatic disease, but questions remain about root causes of this difference and the potential impact on treatment strategies. In this study, we addressed these questions through the development of a mathematical model of PDAC progression that incorporates the major alteration status of specific genes with predictive utility. The model successfully reproduced clinical outcomes regarding metastatic patterns and the genetic alteration status of patients from two independent cohorts from the United States and Japan. Using this model, we defined a candidate predictive signature in patients with low metastatic propensity. If a primary tumor contained a small fraction of cells with *KRAS* and additional alterations to *CDKN2A*, *TP53*, or *SMAD4* genes, the patient was likely to exhibit low metastatic propensity. By using this predictive signature, we computationally simulated a set of clinical trials to model whether this subgroup would benefit from locally intensive therapies such as surgery or radiation therapy. The largest overall survival benefit resulted from complete resection, followed by adjuvant chemoradiation therapy and salvage therapies for isolated recurrence. While requiring prospective validation in a clinical trial, our results suggest a new tool to help personalize care in PDAC patients in seeking the most effective therapeutic modality for each individual.

2. Implementation of Cancer Clinical Sequence

In collaboration with Professor Yoichi Furukawa (Division of Clinical Genome Research, Advanced Clinical Research Center), Professor Arinobu Tojo (Division of Molecular Therapy, Advanced Clinical Research Center), Research Associate Professor Koichiro Yuji (Project Division of International Advanced Medical Research), IMSUT Research Hospital, and Human Genome Center, we have been implementing cancer genomic medicine since 2011. Some reports are published [1,2,5].

Publications

1. Jimbo K, Yokoyama K, Ogawa M, Hirano M, Ochi K, Kobayashi M, Yusa N, Shimizu E, Kawamata T, Yasui H, Ohno N, Yamaguchi R, Imoto S, Furukawa Y, Miyano S, Imai Y, Tojo A. Autologous peripheral blood stem cell transplantation for double-refractory myeloma with K-RAS and N-RAS mutations. *Rinsho Ketsueki.* 58(12): 2380-2385, 2017.
2. Kobayashi M, Yokoyama K, Shimizu E, Yusa N, Ito M, Yamaguchi R, Imoto S, Miyano S, Tojo A. Phenotype-based gene analysis allowed successful diagnosis of X-linked neutropenia associated with a novel WASp mutation. *Ann Hematol.* 97(2): 367-369, 2018.
3. Matsui Y, Niida A, Uchi R, Mimori K, Miyano S, Shimamura T. phyC: Clustering cancer evolutionary trees. *PLoS Comput Biol*. 13(5): e1005509, 2017.
4. Niida A, Nagayama S, Miyano S, Mimori K. Understanding intratumor heterogeneity by combining genome analysis and mathematical modeling. *Cancer Sci.* 2018 Jan 20. doi: 10.1111/cas.13510.
5. Ogawa M, Yokoyama K, Hirano M, Jimbo K, Ochi K, Kawamata T, Ohno N, Shimizu E, Yokoyama N, Yamaguchi R, Imoto S, Uchimaru K, Takahashi N, Miyano S, Imai Y, Tojo A. Different clonal dynamics of chronic myeloid leukaemia between bone marrow and the central nervous system. *Br J Haematol.* 2017 Dec 19. doi: 10.1111/bjh.15065.
6. Park H, Niida A, Imoto S, Miyano S. Interaction-Based Feature Selection for Uncovering Cancer Driver Genes Through Copy Number-Driven Expression Level. *J Comput Biol.* 24(2): 138-152, 2017.
7. Sakai E, Nakayama M, Oshima H, Kouyama Y, Niida A, Fujii S, Ochiai A, Nakayama KI, Mimori K, Suzuki Y, Hong CP, Ock CY, Kim SJ, Oshima M. Combined mutation of Apc, Kras and Tgfbr2 effectively drives metastasis of intestinal cancer. *Cancer Res.* 2017 Dec 27. pii:

canres.3303.2017. doi: 10.1158/0008-5472.CAN-17-3303.
8. Takahashi Y, Sugimachi K, Yamamoto K, Niida A, Shimamura T, Sato T, Watanabe M, Tanaka J, Kudo S, Sugihara K, Hase K, Kusunoki M, Yamada K, Shimada Y, Moriya Y, Suzuki Y, Miyano S, Mori M, Mimori K. Japanese genome-wide association study identifies a significant colorectal cancer susceptibility locus at chromosome 10p14. *Cancer Sci.* 108(11): 2239-2247, 2017.
9. Tominaga K, Shimamura T, Kimura N, Murayama T, Matsubara D, Kanauchi H, Niida A, Shimizu S, Nishioka K, Tsuji EI, Yano M, Sugano S, Shimono Y, Ishii H, Saya H, Mori M, Akashi K, Tada KI, Ogawa T, Tojo A, Miyano S, Gotoh N. Addiction to the IGF2-ID1-IGF2 circuit for maintenance of the breast cancer stem-like cells. *Oncogene.* 36(9): 1276-1286, 2017.
10. Uchi R, Takahashi Y, Niida A, Shimamura T, Hirata H, Sugimachi K, Sawada G, Iwaya T, Kurashige J, Shinden Y, Iguchi T, Eguchi H, Chiba K, Shiraishi Y, Nagae G, Yoshida K, Nagata Y, Haeno H, Yamamoto H, Ishii H, Doki Y, Iinuma H, Sasaki S, Nagayama S, Yamada K, Yachida S, Kato M, Shibata T, Oki E, Saeki H, Shirabe K, Oda Y, Maehara Y, Komune S, Mori M, Suzuki Y, Yamamoto K, Aburatani H, Ogawa S, Miyano S, Mimori K. Correction: Integrated Multiregional Analysis Proposing a New Model of Colorectal Cancer Evolution. *PLoS Genet.* 13(5): e1006798, 2017.
11. Watanabe T, Marotta M, Suzuki R, Diede SJ, Tapscott SJ, Niida A, Chen X, Mouakkad L, Kondratova A, Giuliano AE, Orsulic S, Tanaka H. Impediment of replication forks by long non-coding RNA provokes chromosomal rearrangements by error-prone restart. *Cell Rep.* 21 (8): 2223-2235, 2017.
12. Yamamoto KN, Yachida S, Nakamura A, Niida A, Oshima M, De S, Rosati LM, Herman JM, Iacobuzio-Donahue CA, Haeno H. Personalized management of pancreatic ductal adenocarcinoma patients through computational modeling. *Cancer Res.* 77(12): 3325-3335, 2017.