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
Associate Professor	Sung-Joon Park, Ph.D.	准教授	博士(工学)	朴　　　聖　俊
Assistant Professor	Martin Loza, Ph.D.	助　教	博士(理学)	ローザ　マーティン

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

Professor	Kenta Nakai, Ph.D.	教　授	博士(理学)	中　井　謙　太

The mission of our laboratory is to conduct computational ("in silico") studies on the functional aspects of genome information. At present, we mainly focus on the analysis of regulatory information of gene expression in the non-coding region, using a variety of next generation sequencing (NGS) data. In addition, we are actively collaborating with researchers from various fields.

1. Computational inference of cooperative transcriptional regulators from 3D genomic contacts in germinal center B cells

Sung-Joon Park and Kenta Nakai

During the process of cell differentiation, the higher-order chromosomal structure is intricately remodeled to ensure lineage-specific transcription. In the case of peripheral B-cell differentiation, upon activation by antigen, mature naive B cells (NBs) enter the germinal center reaction along with complex and multifaceted genomic/epigenomic modifications including drastic changes in the 3D chromatin architecture. To understand the impact of chromatin remodeling in germinal center B cells (GCBs) differentiation, we analyzed muti-omics NGS data including Hi-C, Histone ChIP-seq, and RNA-seq with human NB, GCB, and malignant lymphoma cells. We first developed computational models to infer potential transcription factors that bind the promoter regions and long-range contact (LRC) regions. Next, we employed a graph-embedding approach to detect the gene regulatory modules by incorporating cofactors that link promoter-binding TFs and LRC-binding TFs for a certain gene. The in-silico models revealed the distinct transcriptional regulatory roles of promoter-binding and LRC-binding TFs in the context of differentially expressed genes (DEGs) during GCB differentiation. Furthermore, the DEG-TF-Cofactor modules we detected exhibited highly correlated expression patterns in the healthy and malignant B cells. These results provide further explanation for the importance of long-range-contact-mediated transcriptional regulation.

2. Comprehensive Comparison of Gene Expression Diversity among a Variety of Human Stem Cells

Yukiyo Yamatani and Kenta Nakai

Several factors, including tissue origins and culture conditions, affect the gene expression of undifferentiated stem cells. However, understanding the basic identity across different stem cells has not been pursued well despite its importance in stem cell biology. Thus, we aimed to rank the relative importance of multiple factors to gene expression profile among undifferentiated human stem cells by analyzing publicly available RNA-seq datasets. We first conducted batch effect correction to avoid undefined variance in the dataset as possible. Next, we clustered stem cell samples, followed by a dimensionality reduction of their gene expression profile. Then, we highlighted the relative impact of biological and technical factors among undifferentiated stem cell types: a more influence on tissue origins in induced pluripotent stem cells (iPSCs) than in other stem cell types; a stronger impact of culture condition in embryonic stem cells (ESCs) and somatic stem cell types, including mesenchymal stem cells (MSCs) and hematopoietic stem cells (HSCs). In addition, we found that a characteristic gene module, which was enriched in histones, exhibits higher expression across different stem cell types that were annotated by the specific stem cell and growth conditions. This tendency was also observed in mouse undifferentiated stem cell RNA-seq data. We further detected 16 zinc finger family genes that exhibit a co-expression pattern with the histone genes. Altogether, we characterized the major factor, the relative strength of its impacts in each stem cell type, and the impact of culture conditions on histones and zinc finger family genes, which might be highly involved in stem cell identity. Our findings could allow experimental researchers to obtain general insights into stem cell quality, such as the balance of differentiation potentials that undifferentiated stem cells possess.

3. Tc1-Mariner Superfamily of DNA Transposons is Enriched at the Enhancer Boundaries in Ciona

Satoshi Otaki and Kenta Nakai

A long-time stratification of experimental studies on tissue specific enhancers enlisted in the early developmental stage of Ciona allows us to have clues to know different stretch of DNA sequences are involved to cause spaciotemporal gene expression patterns. Less likely found than on mammalian genomes, there are transposable elements interspersed on Ciona genome. There are a growing number of evidence accumulated to suggest some of transposable elements or its remnants reside in host genome playing a role of promoters or enhancers for transcriptional regulation in several model organisms. However, there are now few studies about cis-regulatory elements comprised of transposable elements on Ciona genome. Therefore, we have manually surveyed cis-regulatory regions from literature published from 1997 to 2020. These experimentally validated cis-regulatory regions were mapped on the reference genomes to be modelized as database resources. In our study, though transposable elements are located rarely in exons, one of transposable elements of DNA/TcMar-Tc1 abundantly overlaps with cis-regulatory regions, and we observed an enrichment peak of DNA/TcMar-Tc1 which is a few kb away from neurogenic tissue specific enhancers obtained from construct truncation studies. Furthermore, this transposable element of DNA/TcMar-Tc1 is found to be located just upstream of functional center unit of enhancers. Some of truncation study reveal that this transposable element does not repress aimed gene expression but seems more secure to activate it like a co-working enhancer. Human genome is known to harbor DNA sequence embedded by horizontal transfers of retrovirus genome in ancient time, but many of its sequence are stabilized in the host genome. These retro-elements like LTR elements or its remnants derived from some ERVs work like enhancers in human diseases or neurological disorders, such as multiple sclerosis, amyotrophic lateral sclerosis, and schizophrenia. DNA/TcMar-Tc1 closely located to transcription unit in Ciona has some important role such as demarcation for its boundaries or cis-regulatory elements, but further investigation is required.

4. A variety of nucleotide/amino acid sequence analyses based on deep learning and related methods

Leyi Wei[1] and Kenta Nakai
[1]Shandong University

We have collaborated in a variety of nucleotide/protein sequence analyses. First, we benchmarked 12 available methods for the imputation of scRNA-seq data, based on six simulated and two real datasets. We found that deep learning-based approaches generally exhibit better overall performance than model-based ones. We built an online platform that integrates all available state-of-the-art imputation methods for comparison and visualization analysis. Second, we predicted protein-peptide binding residues via interpretable deep learning; more specifically, we used a BERT-based contrastive learning framework. Since we used a well pre-trained protein language model, we did not need to design effective features for prediction basically but if we integrate the traditional features with our learned features, we could outperform existing methods. Third, we developed a multi-scale deep biological language learning model that enables the interpretable prediction of DNA methylations bed on genomic sequences only. Our model not only outperforms existing methods but also can capture both sequential and functional semantic information from background genomes. Since the model can explain what it learnt, its results will be useful for in-depth analysis of their biological functions. Fourth, by combining all the above works and expanding them, we developed an automated and interpretable platform for biological sequence prediction, functional annotation, and visualization analysis. The platform is accessible as a one-stop-shop web server where researchers are expected to develop a new deep-learning architecture to answer any biological question.

5. Housekeeping enhancers in the human genome

Martin Loza, Alexis Vandenbon[2] and Kenta Nakai
[2]Institute for Life and Medical Sciences, Kyoto University

Enhancers are cis-regulatory elements that regulate cell type-specific gene expression patterns. Recent evidence suggests that a large number of enhancers are active in multiple organs in diverse developmental contexts. However, due to the specificity and complexity of the experiments, the validation of enhancers is not easy to determine in many cell types. To fill this gap, bioinformatics methods have been developed to predict enhancer-gene interactions using various classes of sequencing data, e.g., histone modification and chromatin conformation. In this study, we used the interactions predicted by the ABC method in 50 cell types to characterize the human enhancers. We show that even though most of the predictions are active in only one cell, there are a significant number of conserved elements active among all the 50 cell types. Most of these conserved elements overlap with promoter regions; however, we found around 700 housekeeping enhancers with distinctive characteristics that differentiate them from cell type-specific enhancers. Housekeeping enhancers are rich in G/C nucleotides, and they regulate genes in a longer distance than cell type-specific ones. Besides their low number, as compared with cell type-specific enhancers, housekeeping enhancers regulate around 50% of the protein-coding genes and their distribution correlates across chromosomes. Overall, our work unveils a new type of enhancers conserved in multiple cell types, which will broaden our understanding of the epigenetics behind gene regulation.

6. Integrative Single-cell Analyses of Human Haematopoietic Progenitors Reveals a Putative Dendritic Cell Progenitor in Granulocyte-Monocyte-Dendritic Cell Progenitor

Phit Ling Tan, Florent Ginhoux[3] and Kenta Nakai
[3]Singapore Immunology Network (SIgN), A*STAR

The human dendritic cells (DC) population comprises a heterogeneous family of immune cells, including plasmacytoid DC (pDC) and two subsets of conventional DC (cDC1 and cDC2). Despite the well characterization of mature DC, the origins and differentiation pathways of human DC are still not clearly elucidated. In this study, eight haematopoietic datasets were integrated with Mutual Nearest Neighbors (Haghverdi *et al.*, 2018). A group of bone marrow-derived cells among granulocyte-monocyte-dendritic cell progenitors (GMDP) were found to have the ability to differentiate into cDC via in silico trajectory inference. The group of cells have hallmarks of the DC lineage, including IRF8, CD74, FLT3, and MHC class II expression. A suggested marker panel $CD34^{+}$ $CD123^{int}CD2^{int}CD127^{lo}CD11c^{+}$ was identified using Hypergate (Becht et al., 2019). In the future, we intend to further analyze the group of cells using combinations of flow cytometry, CyTOF and single cell RNA-sequencing, in order to find out its differentiation ability and functional features.

7. Gene Regulatory Network Comparison of Photosensitive Related Cells Between Five Species by Single Cell RNA-seq data

Xin Zeng, Fuki Gyoja[4], Takehiro Kusakabe[4] and Kenta Nakai
[4]Faculty of Science and Engineering, Konan University

The visual system plays an important role in supporting the vertebrates adapted to various environment on the earth. Although the single cell expression profiles of photosensitive related cells from ascidian and vertebrates have been determined, a systematic comparison of developmental mechanisms of homologous cell types between species is needed. Here, we reconstruct the gene regulatory network (GRN) in the photosensitive related cells from five representative species based on gene-gene co-expression correlation by using single-cell RNA-seq data. We compared the GRN for each homologous cell type from these species to identify conserved genetic network. By combining RNA velocity and GRN analysis, we identified the key regulators for retinal cell differentiation in zebrafish and mouse. Altogether, our analysis provides a new understanding of evolutionary molecular mechanisms in photosensitive related cells.

8. Identification of macrophage polarization hysteresis genes using multi-omic datasets

Yubo Zhang, Yutaro Kumagai[5], Sung-Joon Park, Wenbo Yang and Kenta Nakai
[5]Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST)

Macrophages are plastic innate immune cells that can be polarized into classical pro-inflammatory phenotype (M1) and alternative anti-inflammatory phenotype (M2). Accumulating evidence shows that current macrophages state also relies only on its polarization memory (named as macrophage hysteresis). However, the macrophage polarization memory is still not fully investigated. To contribute to this unfulfilled field, we utilized public time-course RNAseq and single cell RNAseq datasets and selected potential memory genes of M1 and M2 phenotypes using traditional machine learning and graph embedding methods. By checking the expression profile of the gene list from a separate single cell RNAseq dataset, we confirmed the functional importance of these genes for macrophage. To further explain the macrophage memory and hysteresis phenomenon, we explored the histone modification among M0, M1, and repolarized M0 macrophages using public ATACseq datasets. We found several TFBS which highly related to inflammatory state are significantly enriched in hysteresis regions. We also applied enhancer prediction using ABC model. These results provided more explanation from epigenetic perspective. Finally, we explore memory genes on patient cancer datasets to further explore the biological meaning of macrophage hysteresis. Overall, our results might improve the progress toward the understanding of macrophage immunity.

9. Spatial Transcriptome Analysis of the Adult Brain of *Ciona intestinalis*

Xin Zeng, Fuki Gyoja[4], Takehiro Kusakabe[4], Yutaka Suzuki[6] and Kenta Nakai
[6]Department of Computational Biology and Medical Science, Graduate School of Frontier Sciences, The University of Tokyo

As one of the closest relatives of vertebrates, the ascidians is a simply but comparable model system to provide insights into the evolutionary process of the chordate nervous system. However, the entire transcriptional profiles of nervous system of adult ascidians have yet to be determined and functionally investigated. Here, we performed 10X Visium spatial transcriptomics system on around 2,000 spots from the adult brain of Ciona. Using unsupervised clustering and Gene Ontology (GO) analysis, we identified four main tissues: body wall muscle, dorsal strand, neural gland, and ciliated groove. Furthermore, we characterized micrometer-scale spatial variable genes and the spatial correlation between genes by a deep generative model and a multivariate normal likelihood ratio test. Together, our analysis sheds the light on spatial gene expression patterns that define the organization of the Ciona adult brain.

10. Constructing a data integration platform for the development of therapeutic agents of COVID-19

Sung-Joon Park, Katsunori Fujiki[7], Satoshi Otaki, Yumiko Imai[8], Katsuhiko Shirahige[7] and Kenta Nakai
[7]Institute for Quantitative Biosciences, The University of Tokyo
[8]National Institutes of Biomedical Innovation, Health and Nutrition (NIBIOHN)

In the medical treatment for patients with respiratory failure including the new coronavirus infection COVID-19, there are still many unexplained factors that influence the effect of therapeutic drugs and vaccines. To contribute to the development of therapeutic drugs for COVID-19, we are developing effective next-generation sequencing (NGS) protocols and a data analysis platform. By collecting bronchoalveolar lavage fluid samples of more than 100 patients with COVID-19 and integrating the information from medical records, we tuned our pipeline for meta-genome analysis and quantified the time-course viral DNA and RNA sequencing data. We found severe bacterial and viral infections known to the involvement of such as myocarditis, pericarditis, and pneumonia. Furthermore, we developed a graph-based machine learning method to predict severity using microbial co-occurrence networks that have been profiled from the patient data, which needs further verification. We believe that the information on bacterial and viral microorganisms significantly present in patients with COVID-19 is a useful resource to establish a therapeutic strategy.

11. Developing an open-access repository for the multi-dimensional genome structure data

Sung-Joon Park, Katsuhiko Shirahige[7], Shoji Takada[9] and Tomoko Nishiyama[10]
[9]Department of Biophysics, Graduate School of Science, Kyoto University
[10]Division of Biological Sciences, Graduate School of Science, Nagoya University

The community-wide effort to characterize 3D genome organization has highlighted the importance of functional linkages between genetic/epigenetic phenomena and physical properties of DNA (e.g., stiffness, torsion, and supercoiling). However, mechanisms underlying the establishment of functional genome structure are still poorly understood, which needs comprehensive and integrative approaches. Here, we are developing a data repository system that facilitates online in-silico analyses, named Genome Modality Suite as a part of the research project Genome Modality. The system has been designed to deal with heterogeneous and multi-dimensional data, such as RNA-seq and ChIP-seq signal tracks (1D data), Hi-C contact matrix (2D data), and XYZ-coordinate structures (3D data). Furthermore, by utilizing PHP, MySQL, and JavaScript libraries, we successfully developed the prototype of a web-based browser to provide seamless access to the 123D data. Our system will accelerate progress toward the understanding of multi-dimensional genome properties.

12. Single-cell transcriptome analysis of ocular-like cell lineages derived from human pluripotent stem cells

Sung-Joon Park, Toru Okubo[11], Yuki Ishikawa[11] and Ryuhei Hayashi[11]
[11]Department of Stem Cells and Applied Medicine, Graduate School of Medicine, Osaka University

The recent cell culturing protocol of the so-called SEAM (self-formed ectodermal autonomous multi-zone), generating 2D eye-like circular colonies from induced pluripotent stem cells (iPSCs), demonstrates that the cells in each zone have different morphologies, e.g., neuroectoderm, ocular-surface ectoderm, and neural crest. To understand the underlying molecular mechanism for the differentiation potency, we are profiling cell populations and their characteristics using time-course scRNA-seq data of SEAM. We confirmed that the undifferentiated iPSCs undergo specific ocular-like cell lineages, including lens cells, retinal cells, and ocular-surface ectoderm progenitors, along with significant marker gene expression. We believe that this single-cell atlas advances studying gene regulation for ocular lineage differentiation.

13. Intercellular crosstalk in adult dental pulp is mediated by heparin-binding growth factors Pleiotrophin and Midkine

Natnicha Jiravejchakul[12,13], Gabriela Abe[14], Martin Loza, Soyoung Park[15], Ponpan Matangkasombut[13], Jun-Ichi Sasaki[16], Satoshi Imazato[14,16], Diego Diez[17] and Daron Standley[12,15]
[12]Department of Genome Informatics, Research Institute for Microbial Diseases, Osaka University
[13]Department of Microbiology, Faculty of Science, Mahidol University
[14]Department of Advanced Functional Materials Science, Osaka University
[15]Department of Systems Immunology, Immunology Frontier Research Institute (IFReC), Osaka University
[16]Department of Biomaterials Science, Osaka University
[17]Quantitative Immunology Research Unit, Immunology Frontier Research Institute (IFReC), Osaka University

The cell heterogeneity and intercommunication in dental pulp (DP) are not well understood. To address this lack of knowledge, we performed an in-silico analysis of publicly available single-cell RNA-seq data from DP. We compared the cell composition in DP against five other reference tissues: blood, bone marrow, adipose tissue, lung, and skin. We identified a DP-specific population of fibroblasts expressing higher levels of heparin-binding growth factors, pleiotrophin (PTN) and midkine (MDK), as compared with fibroblasts from other tissues. We performed cell-cell crosstalk analysis of these DP-specific fibroblasts finding extensive communication with other cell types such as Schwann cells and odontoblasts. Moreover, the analysis revealed that the communication was mostly carried through PTN and MDK receptor binding, which suggests the importance of these molecules in cell proliferation and differentiation in DP. Together, our analysis extends our understanding of cell heterogeneity in DP and their crosstalk communication, which we expect has a potential role in designing and developing therapeutic targeting molecules in dental treatment.

Publication list

Nakai, K,. and Vandenbon,A. (Chapter 2) Higher-order chromatin structure and gene regulation.
(In Chandra Boosani and Ritobrata Goswami eds.) *Epigenetics in Organ Specific Disorders*. pp.11-32, 2022. Academic Press. ISBN: 978-0-12-823931-5

Hayashi, R,. Okubo,T,. Kudo, Y,. Ishikawa, Y,. Imaizumi, T,. Suzuki, K,. Shibata, S,. Katayama, T,. Park, SJ,. Young, RD,. Quantock, AJ. and Nishida, K. Generation of 3D lacrimal gland organoids from human pluripotent stem cells. *Nature* 605:126-131 (2022)

Ding,W,. Nakai, K. and Gong, H. Protein design via deep learning (review). *Brief. Bioinformatics* 23(3) bbac102 (2022)..

Dai, C,. Jiang, Y,. Yin, C,. Su, R,. Zeng, A,. Zou, Q,. Nakai, K. and Wei, L. scIMC: a platform for benchmarking comparison and visualization analysis of scRNA-seq data imputation methods. *Nucl. Acids*

Res., 50(9), 4877-4899 (2022).

Nakai, K. and Wei, L. Recent advances in the prediction of subcellular localization of proteins and related topics. (review) *Front. Bioinformatics*, 2, 910531 (2022).

Wang,R,. Jin, J,. Zou, Q,. Nakai, K. and Wei, L. Predicting protein-peptide binding residues via interpretable deep learning. *Bioinformatics*, 38(13), 3351-3360 (2022).

Kubota, Y,. Fujioka, Y,. Patil, A,. Takagi, Y,. Matsubara, D,. Iijima, M,. Momose, I,. Naka, R,. Nakai, K,. Noda, N. and Takekawa, M. Qualitative differences in disease-associated MEK mutants reveal molecular signatures and aberrant signaling-crosstalk in cancer. *Nature Comm.* 13, 4063 (2022).

Satsu, H,. Gondo, Y,. Shimanaka, H,. Imae, M,. Murakami, S,. Watari, K,. Wakabayashi, S,. Park, SJ,. Nakai, K. and Makoto Shimizu. Signaling pathway of taurine-induced up-regulation of TXNIP. *Metabolites*, 12(7), 636 (2022).

Jin, J,. Yu, Y,. Wang, R,. Zeng, X,. Pang, C,. Jiang, Y,. Li, Z,. Dai, Y,. Su, R,. Zou,Q,. Nakai, K. and Wei, L. iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. *Genome Biology*, 23, 219 (2022)..

Yamatani, Y. and Nakai, K. Comprehensive comparison of gene expression diversity among a variety of human stem cells. *NAR Genomics and Bioinformatics*, 4(4), lqac087 (2022).