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 de Jesus Loza Lopez, Ph.D.	助　教　博士(生命機能学)	ロサロペス　マルティンデヘスス

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

Our laboratory's mission is to conduct computational ("in silico") studies on the functional aspects of genome information. At present, we primarily analyze regulatory information on gene expression in the non-coding region using various next-generation sequencing (NGS) data. We also actively collaborate with researchers from various fields.

1. Epigenetic characterization of housekeeping core promoters in the human genome

Martin Loza, Alexis Vandenbon[1], and Kenta Nakai
[1]Institute for Life and Medical Sciences, Kyoto University

This study investigates housekeeping cis-regulatory elements (HK-CREs) and their roles in human gene regulation and oncogenesis. By integrating multiomics data, we identified approximately 11,000 HK-CREs characterized by distinct epigenetic features, predominantly located in promoter regions. Interestingly, these elements regulate a broader range of targets beyond traditional housekeeping genes. HK-CREs are highly conserved, enriched in unmethylated CpG sites, and engage in extensive long-range interactions with protein-coding genes. Analysis of Hi-C data revealed that HK-CREs associated with zinc finger genes (ZNFs) are enriched in inactive genomic compartments. Despite transcriptional variations potentially linked to heterochromatin marks like H3K9me3, these regions exhibit similar epigenetic profiles and significant promoter–promoter interactions. These findings uncover a previously unrecognized regulatory network, suggesting that HK-CREs serve as fundamental hubs for complex genomic organization and long-range transcriptional control.

2. Identification of potential transcription factors associated with the protection of chromatin loops in CdLS

Martin Loza, Toyonori Sakata[2], Katsuhiko Shirahige[2], and Kenta Nakai
[2]Institute for Quantitative Biosciences, The Univer-

sity of Tokyo

Cornelia de Lange syndrome (CdLS) results from mutations in chromatin organizers like NIPBL, leading to disrupted 3D genome architecture. In this study, we utilized Micro-C data from wild-type and CdLS cell lines to distinguish chromatin loops significantly affected by NIPBL mutations from those that remain stable. We classified these loops by interaction strength and compared their genomic and regulatory landscapes using unsupervised analysis. By integrating transcription factor (TF) binding profiles and epigenetic annotations, we identified specific candidate TFs enriched at unaffected loops. These findings suggest a compensatory mechanism where certain TFs may safeguard loop stability despite impaired NIPBL function. This data-driven framework offers new insights into the molecular resilience of specific chromatin interactions and identifies potential targets for understanding the complex regulatory landscape of CdLS.

3. Bioinformatics analysis of regulatory relationships between long non-coding RNA genes and transcription factor coding genes

Taiju Yamada, Martin Loza, Izuho Hatada[3], and Kenta Nakai
[3]Institute for Molecular and Cellular Regulation, Gunma University

Long non-coding RNAs (lncRNAs) are pivotal regulators of gene expression, yet their systemic relationships with transcription factors (TFs) remain insufficiently characterized. While context-specific lncRNA–TF associations have been reported, a comprehensive analysis of their spatial and transcriptional co-occurrence is lacking. This study investigates the genomic proximity and co-expression landscapes of lncRNAs and TF-encoding genes in humans. By integrating genome-wide annotations with transcriptomic data, we identified consistent spatial clusters and tissue-specific co-expression modules between lncRNAs and TFs. Furthermore, time-series analyses revealed dynamic correlations in expression, suggesting highly coordinated regulatory programs. Our findings provide a systematic overview of lncRNA–TF interdependencies, highlighting a potential regulatory link at both the genomic and transcriptional levels. This work serves as a foundation for elucidating how lncRNAs modulate the core transcriptional machinery across diverse biological contexts.

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

Sung-Joon Park, Katsuhiko Shirahige[2], Shoji Takada[4] and Tomoko Nishiyama[5]
[4]Department of Biophysics, Graduate School of Science, Kyoto University
[5]Division of Biological Sciences, Graduate School of Science, Kyoto University

Three-dimensional genome organization highlights critical functional links between genetic processes and physical DNA properties, such as stiffness and supercoiling. Despite its importance, the mechanisms governing functional genome architecture remain elusive, necessitating integrative analytical approaches. To address this, we developed the Genome Modality Suite (GMS; https://gmsuite.hgc.jp/), a comprehensive data repository and online platform for heterogeneous, multi-dimensional genomic data. GMS integrates diverse data modalities, including RNA-seq/ChIP-seq (1D), Hi-C contact matrices (2D), and spatial genome models (3D), alongside single-cell analysis and protein structure visualization. Built using PHP, MySQL, and JavaScript, the platform is currently live and undergoing continuous refinement to enhance user functionality. By centralizing multi-layered genomic information, GMS aims to provide a robust infrastructure for the community, accelerating discoveries into the multi-dimensional properties that drive genome function.

5. SpliceSelectNet: a hierarchical Transformer-based deep learning model for splice site prediction

Yuna Miyachi and Kenta Nakai

Accurate RNA splicing is vital for gene expression, yet its regulatory mechanisms and the impact of pathogenic mutations remain partially understood. While current splice site prediction tools have advanced, they often struggle with the high computational costs of capturing long-range dependencies (up to 100 kb) and frequently lack interpretability. We present SpliceSelectNet (SSNet), a hierarchical Transformer-based model that predicts splice sites directly from DNA sequences using integrated local and global attention mechanisms. SSNet outperforms state-of-the-art models on the Gencode dataset and achieves superior accuracy for predicting aberrant splicing on the BRCA dataset. Notably, SSNet provides single-nucleotide-level interpretability; its attention maps effectively highlight key cis-regulatory elements, such as exonic and intronic splicing enhancers. By bridging the gap between high-capacity long-range modeling and biological transparency, SSNet offers a powerful framework for elucidating splicing mechanisms and identifying disease-associated disruptions.

6. Deep learning–based prediction of cohesin loader binding from DNA sequence

Yuna Miyachi, Martin Loza, Toyonori Sakata[2],

Kenta Nakai, and Katsuhiko Shirahige[2]

Cohesin is a fundamental regulator of 3D genome organization, yet the mechanisms governing its genomic recruitment remain elusive. Specifically, the sequence-based determinants of the cohesin loader complex NIPBL/MAU2 are poorly understood. We developed a deep learning model to predict NIPBL/MAU2 binding directly from DNA sequences using human ChIP-seq data. Our results demonstrate that integrating epigenetic features, particularly active chromatin marks like H3K27ac, significantly enhances predictive performance, suggesting a functional synergy between cohesin loading and transcriptional activity. Model interpretation via attention analysis and in silico mutagenesis further revealed that distal sequence elements—located several kilobases away—may critically influence NIPBL/MAU2 recruitment. We are currently conducting in vitro experiments to validate these predicted long-range sequence contributions. This framework provides new insights into the sequence-dependent logic underlying cohesin localization and its interplay with the epigenetic landscape.

7. OmniTFT: an EHR-driven multi-target forecasting model for ICU clinical decision support

Wanzhe Xu, Yutong Dai, Yitao Yang, Martin Loza, Weihang Zhang, Yang Cui, Xin Zeng, Sung-Joon Park, and Kenta Nakai

Accurate multivariate time-series forecasting of vital signs and laboratory results is vital for ICU decision support. However, high-frequency noise in vitals and sparse, lagged data in laboratory tests present significant integrative challenges. We propose OmniTFT, a deep learning framework based on the Temporal Fusion Transformer that jointly forecasts these heterogeneous targets. OmniTFT introduces four key strategies: sliding-window equalized sampling for state balancing, frequency-aware embedding shrinkage for stability, hierarchical variable selection, and influence-aligned attention calibration for robustness under physiological shifts. Evaluated on MIMIC-III, MIMIC-IV, and eICU datasets, OmniTFT consistently outperformed existing models. Notably, it improved sepsis mortality prediction by 5% and enabled early AKI detection up to 24 hours in advance. By reducing reliance on target-specific architectures, OmniTFT offers a scalable, interpretable, and generalizable solution for real-time ICU monitoring and early clinical intervention.

8. PhenoNMF: a novel multi-layer matrix factorization framework for age-stratified comprehensive phenotypic similarity analysis

Yutong Dai, Wanzhe Xu, Yitao Yang, Weihang Zhang, Martin Loza, and Kenta Nakai

Electronic health records (EHRs) enable deep phenotyping for risk prediction, yet often lack interpretability and age-specific resolution. Existing single-modality methods often mask physiological variations across life stages, hindering model generalization. To address this, we propose PhenoNMF, a multimodal EHR phenotyping framework based on modified joint non-negative matrix factorization. PhenoNMF identifies sparse phenotypes across diagnoses, laboratory tests, and medications, capturing inter-modal correlations. By reintroducing age through weighted contribution projections, the model isolates multimodality co-occurrences within Common Pattern Models (CPMs) while enabling cross-age comparisons. We introduce the Age Network Coupling Score (ANCS) to rank multimodal triads based on their strength and age specificity. Experimental results demonstrate that PhenoNMF achieves superior clustering cohesion and integration compared to conventional models. Downstream analyses using ANCS quantify age-specific treatment risks and NNT/NNH metrics, providing a robust foundation for precise, age-stratified clinical decision-making.

9. Toward graph-based decoding of tumor evolution: spatial inference of copy number variations

Yujia Zhang[6], Yitao Yang, Yan Kong[7], Bingxu Zhong[6], Kenta Nakai, Hui Lu[6]
[6]Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, China
[7]National Center for Translational Medicine, Shanghai Jiao Tong University, China

Understanding tumor evolution requires mapping spatiotemporal heterogeneity, where copy number variations (CNVs) serve as critical markers. However, current methods often adapt single-cell tools that overlook spatial context, resulting in incomplete clonal architectures. To address this, we developed SCOIGET (Spatial COpy number Inference by Graph on Evolution of Tumor), a framework utilizing graph neural networks with attention layers to infer CNVs by learning spatial neighborhood features from gene expression. SCOIGET significantly reduces error metrics—including MSE and cosine distance—while achieving superior clustering performance over existing methods. Validated on simulated data and patient cohorts, our model accurately depicts the tumor microenvironment across various platforms (10x Visium, Visium HD) and cancer types (colorectal, prostate). By integrating spatial multi-omics features, SCOIGET provides a robust, generalizable solution for decoding tumor evolution, ultimately informing the development of personalized cancer treatment

strategies.

10. Exploration of critical quality attributes (CQAs) for chondrocytes derived from polydactyly patients with guaranteed efficacy

Yasuhisa Ishikawa, Eriko Toyoda[8], Sung-Joon Park, Kenta Nakai, and Masato Sato[8]
[8]Department of Orthopaedic Surgery, Surgical Science, School of Medicine, Tokai University

Osteoarthritis of the knee (OAK) significantly impacts health expectancy and quality of life in aging societies. Currently, no curative treatments exist, making joint replacement inevitable in advanced stages. To address this, Tokai University has pioneered articular cartilage regeneration using chondrocyte sheet transplantation. Clinical studies have demonstrated that autologous and allogeneic sheets effectively repair hyaline cartilage defects, restoring joint function. For clinical implementation, ensuring the uniformity and comparability of allogeneic cell sheets is essential. This study utilizes bioinformatics to identify candidate factors correlating with therapeutic efficacy and to develop an efficacy score assessment model. Establishing these CQAs will facilitate stringent quality control and enable the stable mass production of high-quality cell sheets. Ultimately, these advancements aim to provide a lifelong alternative to artificial joints, allowing patients to preserve their natural joint function.

11. Discovery of broadly neutralizing antibodies with potential therapeutic applications aided by machine learning and artificial intelligence.

Diana Hinojosa-Trujillo[9], Natsrita Piyatida[10], Bowen Liu, Martin Loza, Jesus Hernandez[9], Daron Standley[11], Kenji Mizuguchi[10], and Kenta Nakai
[9]Centro de Investigacion en Alimentacion y Desarrollo, A.C. Hermosillo, Mexico
[10]Institute for Protein Research, Osaka University
[11]WPI Immunology Frontier Research Center, Osaka University

Antibodies are vital therapeutics for cancer and infectious diseases, yet traditional discovery remains labor-intensive and costly. While single-cell RNA sequencing (scRNA-seq) enables rapid identification of antibody sequences from B cells, the vast diversity of these candidates poses a significant bottleneck for functional prioritization. To address this, we are developing an AI-driven bioinformatics pipeline designed to accelerate the discovery of neutralizing antibodies against viral pathogens. By integrating machine learning with high-throughput sequencing data, our framework predicts functional and neutralizing properties, thereby reducing the experimental workload. We are applying this pipeline to diverse viral systems, including SARS-CoV-2 and Zika virus, utilizing public datasets and proprietary libraries from Prof. Hernández's laboratory. This scalable framework aims to streamline antibody development, offering a versatile tool for rapid therapeutic response against both current and emerging infectious diseases.

12. Estimating antibody epitope distance from sequence similarity

Martin Loza, Daron Standley[11], and Kenta Nakai

Quantifying epitope-binding distance is crucial for antibody characterization, yet the scarcity of structural data remains a significant barrier. We developed a two-step computational framework to predict epitope distance using deep mutational scanning (DMS) data and sequence-based features. Initially, XGBoost models were trained on 2,523 non-redundant antibodies to estimate distances for those lacking structural information. Subsequently, we predicted these distances directly from BLOSUM sequence-similarity scores. Our results show that non-linear models, particularly XGBoost, significantly outperformed linear regression. Notably, the simplest data representation yielded the highest accuracy, suggesting that site-level importance is more informative than specific amino acid substitutions. Feature analysis revealed that epitope distance is determined by the collective contributions of multiple CDR sites rather than by a single dominant region. This framework enables robust, sequence-based estimation of epitope distance, bypassing the requirement for high-resolution structural information.

13. Unsupervised and supervised analysis of PB and FNA clinical markers in SLE

Martin Loza, Katia Faliti[12], Melissa Garcia-Vega[12], and Kenta Nakai
[12]Emory University, USA

Systemic lupus erythematosus (SLE) is a heterogeneous autoimmune disease characterized by diverse clinical and serological manifestations. To better understand this complexity, we compared demographic and molecular markers between peripheral blood (PB) and fine-needle aspirate (FNA) samples from patients with SLE. We employed a dual machine-learning approach to dissect the data: unsupervised clustering identified intrinsic patient subgroups with shared characteristics, while supervised models quantified the contribution of specific markers to sample type and clinical status. By integrating these strategies, we aim to elucidate the physiological differences between systemic (PB) and tissue-localized (FNA) measurements. This work identifies clinically relevant patient clusters that are often overlooked by

conventional diagnostic methods, providing a more granular understanding of SLE heterogeneity and potential implications for personalized monitoring and treatment.

14. Temporal and spatial analysis of Xenium data following ear injury in NES and WT mice

Martin Loza, Yitao Yang, Noriko Tokai[13], Mutsuhiro Takekawa[13], and Kenta Nakai
[13]Division of Cell Signaling and Molecular Medicine

Tissue injury triggers complex spatiotemporal shifts in gene expression and cellular composition. This study utilizes Xenium-based spatial transcriptomics to investigate the microenvironment surrounding ear punch injuries in NES-deficient and wild-type (WT) mice. We analyzed samples across a temporal gradient (D0, D1, D3, and D7) to capture the dynamic progression of the wound-healing response. Our bioinformatics pipeline integrates high-resolution spatial clustering, cell-type annotation, and gene set enrichment, specifically focusing on the injury proximity. By synthesizing temporal and condition-specific data, we aim to characterize the evolution of local cellular states and intercellular interactions during the repair process. This work elucidates how molecular signatures and spatial architectures diverge between NES and WT mice, providing a high-resolution map of the factors governing successful versus impaired wound healing.

15. A multi-modal foundation model integrating cancer imaging and molecular data

Mevan Ekanayake[14], Martin Loza, Anna Trigos[14], and Kenta Nakai
[14]Peter McCallum Cancer Center, Melbourne, Australia

Cancer imaging and molecular profiling offer complementary insights into oncology, yet they are typically analyzed in isolation. This project aims to develop a multi-modal foundation model that integrates medical imaging (e.g., PET and CT scans) with high-dimensional genomics and transcriptomics data. By training on diverse multi-modal datasets, the model will capture both shared and modality-specific patterns reflecting tumor biology, heterogeneity, and disease progression. This integrated approach facilitates comparative analyses across various cancer types within a unified computational framework. By bridging the gap between macroscopic imaging features and microscopic molecular profiles, we seek to enhance the characterization of tumor states. Ultimately, this foundation model provides a robust infrastructure for integrative cancer research, supporting more precise diagnostics and a deeper understanding of the complex interactions driving malignancy.

16. Single-cell RNA-seq analysis of cellular communication during ocular lineage specification from iPS cells

Sung-Joon Park, Laura Howard[15], Andrew J Quantock[15] and Ryuhei Hayashi[16]
[15]School of Optometry and Vision Sciences, Cardiff University, UK
[16]Department of Stem Cells and Applied Medicine, Graduate School of Medicine, Osaka University

Human induced pluripotent stem cells (hiPSCs) offer a robust in vitro model for investigating cell fate decisions during early eye development. In this study, we performed single-cell RNA sequencing on hiPSCs undergoing self-organized differentiation into two-dimensional eye-like organoids. Using advanced computational methods, we mapped cellular communication networks and signaling pathways that govern ocular lineage specification. Our analysis identified key transcriptional regulators and signaling dynamics involved in the transition from pluripotency to specialized ocular lineages. Notably, retinoic acid signaling displayed communication patterns consistent with in vivo eye development, alongside conserved tissue-specific transcriptional activity. These findings validate the utility of hiPSC-based systems for dissecting complex intercellular signaling and provide high-resolution insights into the molecular mechanisms guiding early eye development. This framework further supports the development of precise protocols for regenerative medicine and disease modeling.

Publication list

Zhang, Y., Yang, Y., Kong, Y., Zhong, B., Nakai, K., and Lu, H. Toward graph-based decoding of tumor evolution: spatial inference of copy number variations. Diagnostics. 15:3169, 2025.

Khan, A. M., Goto, S., Nakai, K., Wong, L., Kovats, D., Saito, Y., Ikematsu, S., Yamanishi, Y., Wahid, N. S. C., Singh, H., Eranti, P., Chen, Y-P. P., Kim, T-M., Ho, S-Y., Mar, J. C., Iwasaki, W., Valadi, J., Suravajhala, P., Schönbach, C., Tan, T. W., Ranganathan, S., Aoki-Kinoshita, K. F. Establishing the Asia & Pacific Bioinformatics Joint Congress (APBJC): A Historic Milestone in Regional Bioinformatics Collaboration. Brief. Bioinformatics. 26:bbaf678, 2025.

Wang, R., Nakai, K., Wei, L. Prediction of Protein-Peptide Binding Sites Using PepBCL. Methods in Molecular Biology. 2941:269-278, 2025.

Cui, Y., Cui, Y., Wang, R., Zhu, Z., Zeng, X., Nakai, K., Cui, F., Zhang, Z., Shi, H., Chen, Y., Ye, X., Sakurai, T., and Wei, L. DiffusionST: a deep generative diffusion model-based framework for enhancing spatial transcriptomics data quality and identifying spatial domains. Brief. Bioinform. 26:bbaf390, 2025.

Li, J., Sugimoto, E., Yamamoto, K., Dai, Y., Zhang, W., Chang, Y-H., Nakahara, J., Yabushita, T., Kitamura, T., Park, S-J., Nakai, K., Goyama, S. Machine learning-based predictive modeling maximizes the efficacy of mTOR/p53 co-targeting therapy against AML. Cancer Science. 116:3040-3051, 2025.

Cui, Y., Cui, Y., Ding, Y., Nakai, K., Wei, L., Le, Y., Ye, X., and Sakurai, T. OmniClust: A versatile clustering toolkit for single-cell and spatial transcriptomics data. Methods. 238:84-94, 2025.

Park, S-J. and Nakai, K. Genome-Wide Scanning of Gene Expression. Encyclopedia of Bioinformatics and Computational Biology, 2nd Edition. 5:186-196, 2025.

Nakai, K. Prediction of Protein-Binding Sites in DNA Sequences. Encyclopedia of Bioinformatics and Computational Biology, 2nd Edition. 5:239-244, 2025.

Nakai, K., Imai, K., and Wei, L. Prediction of Protein Localisation. Encyclopedia of Bioinformatics and Computational Biology, 2nd Edition. 5:119-126, 2025.

Yang, J., Sato, K., Loza, M., Park, S-J., and Nakai, K. RNA secondary structure prediction by conducting multi-class classifications. Computational and Structural Biotechnology Journal. 27:1449-1459, 2025.

Zeng, X., Liu, S., Liu, B., Zhang, W., Xu, W., Toriumi, F. and Nakai, K. Gene2role: a role-based gene embedding method for comparative analysis of signed gene regulatory networks. BMC Bioinformatics. 26:134, 2025.

Yang, Y., Cui, Y., Zeng, X., Zhang, Y., Loza, M., Park, S-J., and Nakai, K. STAIG: spatial transcriptomics analysis via image-aided graph contrastive learning for domain exploration and alignment-free integration. Nature Comm. 16:1067, 2025.

Zhang, Y., Yang, W., Kumagai, Y., Martin, L., Yang, Y., Park, S-J. and Nakai, K. In-silico analysis revealed Marco (SR-A6) and Abca1/2 as potential regulators of lipid metabolism in M1 macrophage hysteresis. International Journal of Molecular Sciences. 26(1):111, 2025.

Howard, L., Ishikawa, Y., Kamuro, R., Katayama, T., Bains, K. K., Hill, M. J., Blake, D. J., Park, S-J., Hayashi, R., Quantock, A. J., Nishida, K. Single-Cell RNA-Seq Reveals Conserved Cellular Communication Mechanisms Governing Ocular Lineage Specification from Human iPS Cells. Cells. accepted, 2025.