

## Human Genome Center

# Division of Digital Genomics

## デジタル・ゲノミクス分野

| Professor

Natsuhiko Kumasaka, Ph.D.

| 教授

博士(理学)

熊坂夏彦

*A genome-wide association study (GWAS) is a powerful approach for identifying genetic variants and related genes involved in the molecular mechanisms of common complex traits, such as diabetes and human height. As of April 2024, the GWAS Catalogue reports 691,532 genetic associations discovered for 36,643 common complex traits. The Division of Digital Genomics aims to identify these genetic associations of common complex traits and uncover their molecular mechanisms using cutting-edge molecular biology assays and integrated mathematical and statistical approaches.*

### 1. Discovery of genetic determinants for child health and development

**Natsuhiko Kumasaka**

Understanding the influence of both genetics and environment on human health, especially early in life, is essential for shaping long-term health. Here, I utilize a population-based prospective birth cohort, the Japan Environment and Children's Study (JECS), to conduct a large-scale genetic study using questionnaire surveys and biological and physical measurements collected from both parents and their children since the participant mothers were pregnant.

JECS is a large-scale birth cohort study established by the Ministry of the Environment, Government of Japan, to evaluate the effects of environmental chemicals on children's health and development. Over 100,000 pregnant women were enrolled at 15 geographically different regional centers across Japan, representing the comprehensive genetic diversity of the Japanese population. Detailed data from questionnaires, biological and physical measurements have been collected with additional surveys every six months still being conducted on average 70% of the child participants until the age of 4.

My role is to conduct genome-wide association studies using the various child health and developmental outcomes and to make the results available to the public. As of December 2024, genome-wide genotyping analyses have been performed on 80,638 child participants with parental consent and sufficient DNA from cord blood samples. Systematic genome-wide association studies of 1,163 child health and developmental traits (including, for example, food allergy or ASQ-3 developmental screening) and parental environmental exposure traits identified 4,985 common genomic loci, of which a part of loci represented novel associations not previously reported. The results have been tailored as the flagship publication, entitled 'Genome-wide association study on longitudinal and cross-sectional traits of child health and development in a Japanese population', which is currently in the process of paper submission and will appear on bioRxiv in early 2025 prior to full peer review.

### 2. Development of a novel in-vitro approach to validate and elucidate underlying molecular mechanisms of genetic associations discovered through GWAS

Yuji Miyatake, Kotoe Katayama<sup>1</sup>, Seiya Imoto<sup>1</sup>, Ka-

zuaki Yokoyama<sup>2</sup>, Yasuhito Nannya<sup>2</sup>, Atsushi Fukuda<sup>3</sup>, Natsuhiko Kumasaka

<sup>1</sup>Division of Health Medical Intelligence, Human Genome Center, IMSUT, <sup>2</sup>Department of Hematology/Oncology, IMSUT, <sup>3</sup>School of Medicine, Tokai University

Although Genome-wide association studies (GWAS) have identified hundreds of thousands of genetic associations in common complex traits, more than 90% of genetic variants discovered by GWAS (referred to as GWAS variants) are located in non-coding regions. This poses a significant challenge in our efforts to identify putative causal variants and functional genes involved in a regulatory cascade of disease onset and progression. In addition, the target cell type(s) and cellular states in which these GWAS variants affect gene expression often remain unknown, limiting our ability to identify molecular mechanisms of GWAS susceptibility loci in follow-up studies.

Expression quantitative trait locus (eQTL) mapping is a powerful approach to gain insight into the role of non-coding variants in gene regulation. It allows us to discover genes that are regulated by these GWAS variants and helps elucidate downstream consequences. In addition, the recent advances in single-cell genomics allow us to identify target cell types and cellular states in which GWAS variants modulate gene expression through eQTLs. However, the study of eQTLs can often prove to be cost-ineffective and labor-intensive, especially when sample collection is challenging, such as in the case of *in vivo* brain samples collected from hundreds of patients undergoing neurosurgery, or in the case of differentiated neurons derived from hundreds of human pluripotent stem cell (hPSC) lines.

Recently, an alternative approach combining single-cell RNA-seq (scRNA-seq) with massively parallel CRISPR interference (CRISPRi) screening has been proposed to map eQTLs. This approach relies on CRISPR-mediated perturbations instead of natural genetic variation and is theoretically feasible from a single donor's sample. However, even though this approach significantly reduces the cell culture and experimental burden, it has not yet been applied to any brain cell type implicated in neuropsychiatric and neurodegenerative diseases.

Indeed, a combination of a flexible *in vitro* system and a robust *in silico* approach is lacking. Although the use of hPSC is a valuable tool to generate different types of mature cells, it is not trivial to maintain a sufficient gRNA repertoire through cell expansion and differentiation processes due to the selection pressure of specific CRISPR-mediated perturbations (*i.e.*, only cells with specific gRNA expand faster and take over other cells). From a data analysis perspective, the

identification of dynamic genetic effects that manifest only during specific phases of cellular transition has historically been challenging, due to the absence of robust machine learning approaches until very recently.

We combine the unique expertise of machine learning/bioinformatics and stem cell biology/gene editing to develop a novel CRISPR perturbation system based on midbrain organoids established from hPSC lines, coupled with a state-of-the-art machine learning technique using Gaussian processes. We have already established a comprehensive computational approach, GASPACHO (GAUSSian Processes for Association mapping leveraging Cell Heterogeneity), for mapping eQTLs along dynamic cellular states, which can be readily applicable to CRISPR-mediated eQTL mapping as if presence/absence of a gRNA in a cell were different genotypes at a natural genetic variant. We also have established several hPSC lines for CRISPR screening with inducible expression of dCas9-KRAB or dCas9-P300 by doxycycline (Dox), which ensures a more stable cell culture and allows to introduce perturbations at any desired time point of cell differentiation. In addition, we have extensive experience in two-dimensional neuronal differentiation and development of brain organoids.

As of December 2024, we have established both the Dox-inducible CRISPRi and CRISPRa iPS cell lines. Using the CRISPRi line, we developed a mid-brain organoid suitable for studying Parkinson's disease (PD) GWAS loci. We have sequenced approximately 160K cells at the first two different time points of organoid development. This year we plan to sequence a further 160K cells at the two later time points as well as 4 different time points using the CRISPRa line to fully investigate the genetic role of PD-associated variants in the non-coding gene regulatory mechanism.

### 3. Development of novel statistical approaches to map genetic associations using Gaussian Processes

Yuji Miyatake and Natsuhiko Kumasaka

The Gaussian Process (GP) is a powerful approach for modelling non-linear phenomena in scientific fields such as genomics and genetics. This project focuses on the use of GPs for genetic association mapping. The aim is to identify genetic variants that affect gene regulation across continuous cellular states at the molecular level, and disease susceptibility over time and space at the population level. We are currently developing a robust and sensitive method for mapping dynamic genetic effects based on a quasi-Poisson generalized linear mixed model. This method can be applied to different outcomes, includ-

ing non-Gaussian outcomes, to estimate latent factors embedded in a data matrix and map genetic associations with appropriate statistical calibration.

### Publications

JECS Genetics Research Group. Genome-wide association study on longitudinal and cross-sectional traits of child health and development in a Japanese population. In preparation.

Kumasaka N. Genetic Association Mapping Leveraging Gaussian Processes. **J Hum Genet.** 69: 505-510, 2024.