インシリコ解析ユニット

補助事業代表者（所属・職名）
Daron M. Standley（大阪大学 微生物病研究所 ゲノム情報解析分野・教授）

分子設計、超分子モデリング、シミュレーションを用いた バイオマーカーの探索および創薬技術支援

補助事業分担者（所属・職名）
Soyoung Park（東京科学大学 地球生命研究所・准教授）

高度化における研究背景・目的

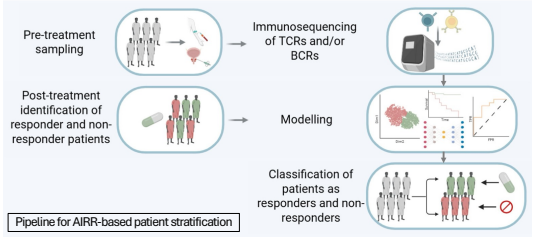
In life science research, there is a growing need for advanced *In silico* analysis platforms capable of analyzing large-scale and diverse datasets with high accuracy and efficiency and translating the resulting information into biological insights. This project aimed to advance methods for large-scale data analysis, machine learning, database construction, annotation, and prediction using life science data, including sequence, genomic, and omics data, and to evaluate their utility through applications in life science research.

高度化の成果

A. AI・大規模配列解析

1. AIRR-based stratification of immunotherapy responses

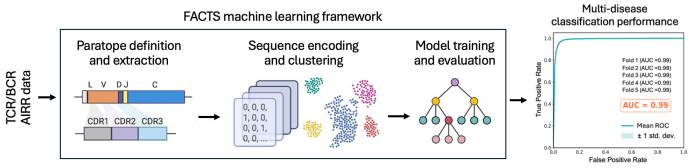
We reviewed current approaches for using Adaptive Immune Receptor Repertoire (AIRR) to stratify immunotherapy responses, including the conventional repertoire metrics and the sequence-based machine-learning, and discussed key requirements for standardization, data sharing, and clinical translation. (Biró et al., *Trends Mol Med*, 2026)



Pipeline for AIRR-based patient stratification

2. Scalable adaptive immune repertoire analysis

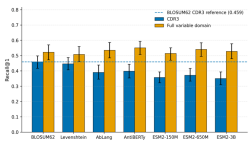
We developed FACTS, a scalable classification framework for adaptive immune repertoire that combines paratope sequence representation, rapid clustering, and machine learning. FACTS enabled multi-disease classification of around 81 million TCR sequences from 1,421 donors, achieving a mean multiclass ROC-AUC of approximately 0.99, at mean accuracy of 81.77%. (Jiravejchakul et al., *bioRxiv*, 2025)



FACTS framework and classification performance on public TCR dataset from 1,421 samples covering 15 diseases and 1 healthy cohort

3. Evaluation of protein language models for immune receptor analysis

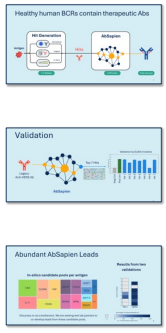
We systematically benchmarked protein language models (PLMs) against conventional sequence-alignment approaches for adaptive immune receptor retrieval. The analysis revealed that much of the apparent PLM advantage for full-length receptor sequences can arise from conserved germline/framework information, highlighting the importance of appropriate sequence representation and dataset splitting. (Wang et al., *bioRxiv*, 2026)



BCR retrieval performance using alignment methods (BLOSUM62, Levenshtein), antibody-specific PLMs (AbLang, AntiBERTy), and ESM2 models, comparing CDR3-only (blue) and full variable-domain (orange) representations (all-CDR paratope input).

4. Identification of therapeutic antibodies in healthy BCR Repertoires

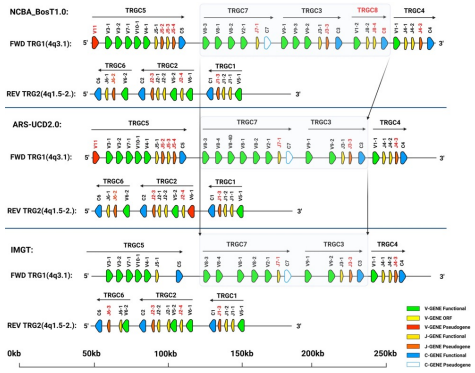
We have developed *AbSapien*, a method of identifying functionally related antibodies across donors and even across species. In our first test, 40% of *AbSapien* hits bound their predicted targets based on ELISA assays. We only tested 20 such hits, but similar predicted antibodies were found found many known therapeutic antibodies.



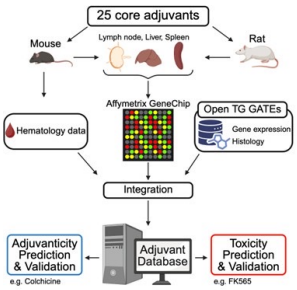
B. ゲノム・データ基盤

1. Improved annotation of immune receptor loci

We developed and applied a comparative genome-alignment strategy to improve immune receptor gene annotation. Application to the bovine TRG locus identified previously unannotated TRG genes and produced an updated annotation by integrating information across multiple genome assemblies. (Zhou et al, *BMC genomics*, 2025)



Cow TRG locus across three genome assemblies. Newly identified genes are shown in red, and shaded regions highlight differences among assemblies.



2. Development of an adjuvant transcriptomics database and prediction framework

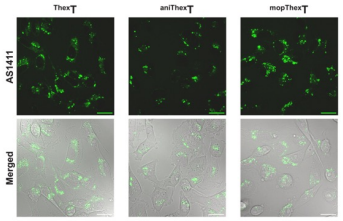
Through collaboration, we established an Adjuvant Database (ADB) containing standardized transcriptomic profiles of 25 adjuvants. Integration with toxicity datasets and machine-learning approaches enabled prediction and experimental evaluation of both adjuvant activity and toxicity. (Natsume-Kitatani et al., *Cell chemical biology*, 2025)

Adjuvant Database framework integrating multi-omics and toxicity data for prediction of adjuvanticity and toxicity.

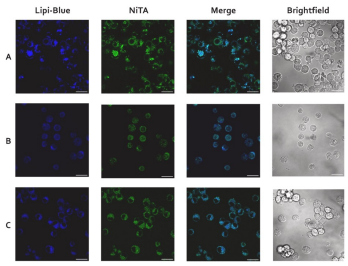
C. 分子解析・可視化技術

1. Development of microenvironment-sensitive fluorescent nucleosides

Through collaborative research, we developed chemically tunable ThexT-derived fluorescent nucleosides whose fluorescence responds to the local molecular environment. The probes can be incorporated into DNA, report G-quadruplex/nucleolin interactions, and enable live-cell imaging when incorporated into the AS1411 aptamer. (Kumagai et al., *Sens Actuators B Chem.* 2025)



Confocal microscopic images of living MDA-MB-231 cells treated with 10 μM of ThexT derivative-modified AS1411 for 3–4 h.



Live-cell imaging with co-staining of Lipi-Blue and NiTA demonstrates the colocalization of lipid droplets in (A) Jurkat cells, (B) LK35.2 cells, and (C) RAW 264.7 cells

2. Development of an environment-sensitive lipid-droplet probe

Through collaborative research, we developed NiTA, a compact environment-sensitive fluorescent probe for lipid droplets. NiTA enabled visualization of lipid droplets in multiple immune-cell types and monitoring of changes in their local lipid environment, including applications in living medaka. (Wee et al., *Anal Chim Acta*, 2025)

今後の発展

We will further advance AI and machine-learning approaches for large-scale life science data, with a particular focus on integrating diverse data types to generate new biological insights and improve predictive accuracy and interpretability. We will also strengthen collaborations with experimental and clinical researchers, expand these approaches to a broader range of life science applications, and develop reusable analytical frameworks and data resources that can be shared across research communities.