No.	K25-2170	
Project Title	Large-scale gene regulatory network estimation from single-cell transcriptome based on high-performance computing	
Principal Investigator	Marco Scutari (Dalle Molle Institute for Artificial Intelligence / University of Applied Sciences and Arts of Southern Switzerland (SUPSI)・Senior Researcher (equivalent to Associate Professor))	
Project Member(s)	IMSUT Host Researcher	Seiya Imoto (The Institute OF Medical Science, The University Of Tokyo・Professor)
	Project Member (s)	Noriaki Sato (Division of Health Medical Intelligence・Assistant Professor)
	Project Member (s)	Kotoe Katayama (Laboratory of Sequence Analysis・Associate Professor)
	Project Member (s)	Rui Yamaguchi (Aichi Cancer Center・Division Head)
	Project Member (s)	Yoshinori Tamada (Graduate School of Medicine, Hirosaki University・Professor)
	Project Member (s)	Shuichi Kawano (Faculty of Mathematics, Kyushu University・Professor)

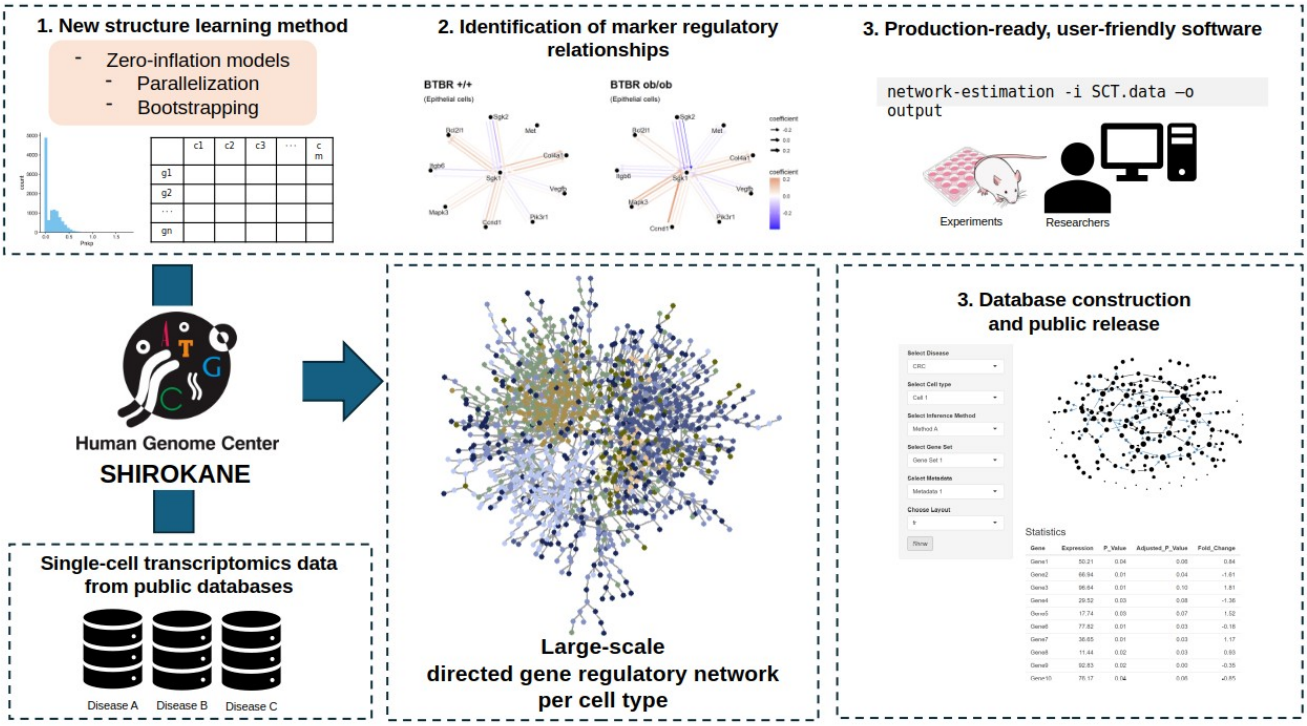
(Form 2)

Principal Investigator
Marco Scutari

IMSUT International Joint Usage/Research Center Project <International>
Joint Research Report (Annual/Project Completion)

Annual Report

Report



Single-cell transcriptomics (SCT) analysis has elucidated detailed mechanisms of disease in medicine by analysing molecular transcripts at the single-cell level. The aim of this project (”Large-scale gene regulatory network estimation from single-cell transcriptome based on high-performance computing”) is to develop inference methods for gene regulatory networks (GRNs) based on Bayesian Networks and tailored to SCT data. Our goals are:

1. Developing GRN estimation methods explicitly tailored to SCT data, from both a statistical (modelling of zero-inflated distributions) and a computational perspectives (scalability beyond a few tens of genes). Incorporate prior information available from annotation databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Dorothea, to improve network reconstruction accuracy.
2. Developing production-ready, user-friendly software for large-scale GRN estimation.
3. Building a public database of directed networks whose arcs represent the interplay between genes causative for diseases.

In FY2025, we have made significant progress towards these goals.

- We developed scstruc, a production-ready, user-friendly R package for large-scale GRN estimation (https://github.com/noriakis/scstruc). We described it in a paper which is currently undergoing the second round of review at Cell Reports Methods (doi:10.48550/arXiv.2511.12805).

- We defined the first causal metric (sSID) to evaluate the effects of molecular interventions on GRNs, using annotation databases as a reference and incorporating the latest research on causal discovery and inference.
    - The full paper describing it has been submitted to Bioinformatics, where it is currently scheduled for the second round of review (doi:10.48550/arXiv.2511.12805).
- We developed a new, differentiable algorithm for GRN discovery that scales to a few hundred genes and outperforms general-purpose Bayesian network approaches in both speed and accuracy.

◦ The paper describing this work (doi:10.48550/arXiv.2512.16233) was presented at the 5th AAAI Workshop on Graphs and More Complex Structures for Learning and Reasoning, colocated with AAAI 2026 and the 5th conference on Causal Learning and Reasoning (CLeaR 2026). It will be published as a full paper in the dedicated volume of the Proceedings of Machine Learning Research (PMLR).

- We are developing a novel Bayesian approach to incorporate the causal metric above (sSID) into GRN reconstruction, building on previous work by Prof. Seiya Imoto on energy-based prior distributions. A preliminary version of this work has already been accepted for presentation at two conferences in FY2026:
    - The 18th World Meeting of the International Society for Bayesian Analysis (ISBA), Nagoya (June 28–July 3, 2026).
    - The 35th European Meeting of Statisticians (EMS), Lugano (August 24–28, 2026).

In addition to these specific contributions, this project has enabled the efficient production-ready implementation of many baseline techniques for reconstructing and evaluating GRNs in the bnlearn R package (https://www.bnlearn.com), which is the leading open-source software for Bayesian networks in statistical machine learning. These include, for instance, the Direct LiNGAM algorithm (doi:10.48550/arXiv.1101.2489), the Structural Interventional Distance (SID, doi:10.1162/NECO_a_00708) and zero-inflated DAG models (https://jmlr.org/papers/v24/21-1476.html). These will benefit the community at large as they can be reused out of the box for other research.

This project has also established long-standing collaborations between the institutions involved on the theme of complex networks, spanning medicine, statistics, biostatistics and bioinformatics. These collaborations will continue in FY 2026, when further research visits are planned throughout the year.