Global Indexes for Genome Association Studies in Microbial Populations

Antoine Limasset

01/12/2023

Context

The project "Global Indexes for Genome Association Studies" is set against the backdrop of significant advancements in statistical genetics, particularly in human populations. The field has made considerable strides in elucidating the relationship between genetic variations and phenotypic traits. This has been primarily facilitated by the reduction in sequencing costs and the development of sophisticated computational methodologies for identifying genetic variants and associating them with phenotypes, notably in genome-wide association studies (GWAS). Despite these advances in human genetics, the application of GWAS in microbial genomics has not kept pace. This lag is intriguing, considering the smaller genome size of microbial entities, which would theoretically simplify analyses. The project aims to understand and address this paradox, focusing on the unique features of microbial genomes that challenge the direct application of human-genetic GWAS methods.

Objectives

This project endeavors to develop and refine GWAS tools specifically tailored to the unique structure of microbial genomes. The primary challenge is the genomic plasticity caused by horizontal gene transfer in microbial entities, leading to considerable genetic variation between isolates. The project hypothesizes that encoding genetic variants as k-mers provides an innovative solution to capture genetic variations effectively. However, the challenge lies in interpreting these k-mers and their causal relationship to phenotypes. The project aims to develop a new suite of microbial GWAS tools that can handle future large-scale genomic datasets and utilize the growing body of publicly available genomic data for variant annotation. This will facilitate a deeper understanding of microbial phenotypes' genetic determinants, addressing a significant gap in current GWAS methodologies.

The project is structured into five work packages, each with specific goals and methodologies:

- WP1: Scalable Structure Design for Genetic Variant Indexing Developing a novel data structure to efficiently index large amounts of k-mers from large gene cluster collection [2, 1].
- WP2: Ultra-Sensitive Variant Association Contextualization

  Innovating graph exploration algorithms to extend the context of genetic variants and enhance interpretability [3].
- WP3: Enhanced Annotation of Genetically Significant Variants

  Leveraging publicly available genomic sequences to annotate genetic variants and assess their phenotypic implications[5].
- WP4: Benchmarking and Validation

  Demonstrating the efficacy of developed methodologies by reanalyzing existing microbial GWAS datasets [4].
- WP5: Management, Coordination, and Dissemination

1 Collaboration

The project brings together experts from different institutions across France and Germany, including ULille(Antoine Limasset), CNRS(Laurent Jacob), and Twincore(Marco Galardini).

2 Impact, Benefits, and Scientific Significance:

The project aims to change the field of microbial GWAS by providing scalable, efficient tools for large-scale genetic studies. By improving the interpretability of GWAS results, the project aims to make these analyses accessible to a wider scientific audience, fostering greater inclusivity and democratizing research across diverse regions. The project also positions itself as a catalyst for future innovations in the field, potentially influencing related applications in genomics and beyond. The collaboration fosters a rich exchange of ideas and expertise, providing a dynamic learning environment for the involved researchers and students.

References

[1] Clément Agret, Bastien Cazaux, and Antoine Limasset. Toward optimal fingerprint indexing for large scale genomics. bioRxiv, 2021.

[2] Camille Marchet and Antoine Limasset. Scalable sequence database search using partitioned aggregated bloom comb-trees. bioRxiv, 2022.

[3] Christopher Quince, Sergey Nurk, Sebastien Raguideau, Robert James, Orkun S Soyer, J Kimberly Summers, Antoine Limasset, A Murat Eren, Rayan Chikhi, and Aaron E Darling. Strong: metagenomics strain resolution on assembly graphs. Genome biology, 22:1–34, 2021.

[4] Hector Roux de Bézieux, Leandro Lima, Fanny Perraudeau, Arnaud Mary, Sandrine Dudoit, and Laurent Jacob. Caldera: Finding all significant de bruijn subgraphs for bacterial gwas. Bioinformatics, 38(Supplement_1):i36–i44, 2022.

[5] Hannes Sommer, Dilfuza Djamalova, and Marco Galardini. Reduced ambiguity and improved interpretability of bacterial genome-wide associations using gene-cluster-centric k-mers. Microbial Genomics, 9(11):001129, 2023.