Title: Improving genotype imputation and unveiling host genetics-microbiome interactions: A study on accuracy and reproducibility in the Japanese population

Author: Ortega Reyes, David
Affiliation: Department of Genetics, School of Life Science, The Graduate University for Advanced Studies, SOKENDAI

Introduction

Genome-wide association studies (GWAS) have significantly advanced our understanding of the genetic foundations of complex traits and diseases. By scanning the genome for statistical associations between genetic markers and traits of interest, GWAS have facilitated the discovery of thousands of genetic loci influencing complex traits. However, the success of GWAS is linked to the quality of genotype imputation, a statistical method that predicts unobserved genotypes in a study sample using a reference panel of haplotypes. Imputation increases genome-wide marker density, enhances the power to detect genetic associations, and allows fine-mapping of causal variants. The accuracy of imputation relies on the size and diversity of the reference panel, as well as the genetic similarity between the study population and the reference panel.

The Japanese population presents unique challenges in genotype imputation due to its distinct genetic architecture shaped by historical isolation. Global reference panels like the 1000 Genomes Project (1KG) and the Haplotype Reference Consortium (HRC), while including some East Asian populations, are predominantly composed of individuals of European ancestry. Consequently, imputation of rare and low-frequency variants in Japanese individuals is often inaccurate, reducing GWAS power and potentially leading to false negatives. Developing population-specific reference panels is essential to improve imputation accuracy and ensure reliable identification of trait-associated variants in underrepresented populations.

Simultaneously, the interplay between host genetics and gut microbiome composition has emerged as a critical area of research. The gut microbiome influences various physiological processes, including metabolism, immune function, and neurological development. Host genetics contribute significantly to shaping microbiome composition, as evidenced by heritability estimates and identification of quantitative trait loci influencing microbial taxa. However, studies exploring host-microbiome associations often face challenges in reproducibility due to variations in study design, sample sizes, sequencing methodologies, and data processing pipelines.

This thesis aims to address two critical areas in genomic research within the Japanese population:

1. Improving imputation quality for GWAS: To develop and evaluate population-specific reference panels by augmenting the 1KG reference panel with Japanese whole-genome sequencing (WGS) data. The goal is to enhance imputation accuracy, particularly for

rare variants, and to demonstrate the utility of these panels in identifying novel genetic associations with serum uric acid and total cholesterol levels.

2. Exploring the interplay between host genetics and gut microbiome composition: To investigate genetic variants associated with gut microbiome composition in the Japanese population through GWAS. Additionally, to assess the reproducibility of findings across studies and examine methodological factors influencing results.

Materials and methods

Part I: Improving imputation quality for GWAS

Study population and WGS data

This study used data from the Biobank Japan (BBJ) project, a large-scale, hospital-based registry collecting genomic and health-related data from approximately 200,000 Japanese individuals. A subset of 7,517 individuals was selected for WGS, recruited from seven geographical regions across Japan to capture genetic diversity. Sequencing was performed at varying depths:

- 1,502 individuals at 30× depth: High-depth sequencing for accurate detection of both common and rare variants.
- 1,786 individuals at 15× depth: Medium-depth sequencing balancing accuracy and sample size.
- 4,229 individuals at 3× depth: Low-depth sequencing to increase haplotype diversity.

Development of population-specific reference panels

Four reference panels were constructed by augmenting the 1KG reference panel with BBJ WGS data:

1. 1KG+1K Panel: Incorporating 1,037 individuals at 30× depth.
2. 1KG+3K Panel: Including 3,256 individuals at 30× and 15× depths.
3. 1KG+4K Panel: Adding 4,216 individuals at 3× depth.
4. 1KG+7K Panel: Combining all 7,472 individuals.

Data processing involved quality control, phasing with SHAPEIT5, and reciprocal imputation using Minimac4 to harmonize the panels.

Evaluation of imputation quality

A leave-one-out cross-validation was performed by excluding 200 high-depth WGS samples for testing. Imputation was conducted on chromosomes 5 and 19 using Minimac4. Imputation quality metrics included aggregated r^2, concordance rate, precision, recall, and Imputation Quality Score (IQS).

GWAS for serum uric acid and total cholesterol

GWAS were conducted using imputed genotypes from approximately 180,000 BBJ participants. Phenotypes included serum uric acid (104,174 individuals) and total cholesterol levels (122,407 individuals). Linear regression models adjusted for age, sex, principal components, and disease status were employed. Results from the 1KG+7K and TOPMed panels were compared.

Part II: Host genetics and gut microbiome composition

Study population and sample collection

A cohort of 306 Japanese individuals aged 20–75 years was recruited during health check-ups at Tokyo University Hospital. Participants were categorized as normal health, obese, or prediabetic based on clinical criteria. Exclusion criteria included prior diagnosis of diabetes, recent antibiotic use, and significant weight loss. Fecal and blood samples were collected following standardized protocols.

Microbiome data generation

Two sequencing methodologies were utilized:

- 16S rRNA Gene Sequencing: Targeting the V1-V2 regions to identify bacterial genera.
- Shotgun Metagenomic Sequencing: Providing species-level resolution and functional profiling.

Taxa and pathways were filtered based on prevalence and abundance to focus on biologically relevant entities.

GWAS of microbiome composition

GWAS were performed between host genetic variants and the inverse-normal-transformed relative abundances of gut microbial taxa and pathways. Approximately 13 million high-quality variants were analyzed using linear regression models adjusted for covariates, including age, sex, sequencing batch, clinical group, and principal components.

Comparative methodological analysis

To assess the impact of methodological differences, a comparative analysis was conducted using two sample processing and DNA extraction methods: one replicating Ishida et al.'s protocol and the other following the current study's protocol.

Results

Part I: Imputation quality improvement

Imputation quality assessment

The 1KG+7K panel outperformed other panels and the TOPMed panel, particularly for rare variants (MAF <5%). For variants with MAF between 1–5%, the aggregated r^2 increased from 0.865 (TOPMed) to 0.919 (1KG+7K). The percentage of variants with concordance r^2 >0.9 was higher with the 1KG+7K panel (75.8%) compared to TOPMed (69.2%). The IQS was consistently higher for the 1KG+7K panel across all MAF bins.

GWAS findings

- Serum uric acid levels: Eleven loci reached genome-wide significance, with five novel loci uniquely identified using the 1KG+7K panel. Notably, variants in *SLC22A13* (rs117371763) and *NRG4* (rs12595289) were associated with decreased uric acid levels.
- Total cholesterol levels: Four potentially novel loci were identified, including a significant rare variant near the *PPIA*/*H2AZ2* genes (rs899749693, MAF = 0.23%). These loci were absent in the TOPMed-imputed results due to poor imputation quality.

Part II: Host genetics and microbiome composition

Genetic associations with microbial taxa

GWAS identified:

- Genus level: Seven significant loci associated with genera such as *Bifidobacterium* and *Blautia*.
- Species level: Thirteen significant loci, including associations with *Bifidobacterium bifidum* and *Bifidobacterium animalis*.
- Bacterial pathways: Eight significant loci associated with metabolic pathways.

Applying Bonferroni correction, one locus at the genus level and two at the species level remained significant. No significant associations were found at the phylum level.

Lack of reproducibility across studies

Comparison with four recent host-microbiome GWAS revealed minimal overlap. Despite analyzing similar numbers of taxa and identifying comparable numbers of significant loci, direct replication of specific SNP-taxon associations was absent across studies. Methodological differences, sample sizes, and population-specific factors likely contributed to inconsistencies.

Impact of methodological differences

The comparative analysis demonstrated significant differences in microbiome composition due to sample preservation and DNA extraction methods. Variations

particularly affected the relative abundances of *Bacteroides* and *Bifidobacterium* genera. This emphasizes that methodological differences can significantly influence the observed composition of the gut microbiome and contribute to inconsistencies between studies.

Discussion

Part I: Interpretation and implications

Integrating Japanese WGS data improved imputation accuracy by increasing sample size and capturing population-specific alleles. The enhanced imputation facilitated the discovery of novel genetic associations, contributing to a better understanding of complex traits. Improved imputation enables the identification of rare variants influencing disease susceptibility, informing personalized medicine approaches.

Part II: Host-microbiome interactions

The GWAS results support a genetic influence on microbiome composition. The associations identified suggest that host genetics contribute to shaping specific components of the gut microbiome. However, the lack of reproducibility underscores the complexity of these interactions and the impact of methodological differences. Standardization is crucial for reproducibility. Variations in sample processing and DNA extraction significantly affect microbiome profiles, hindering cross-study comparisons.

Functional links between genetic variants and microbiome composition

We explored potential functional links between significant host genetic variants and gut microbiome composition. For example, the association between rs79675168-A and *Bifidobacterium bifidum* abundance suggests a role of the gene *LIMA1* in cholesterol metabolism affecting microbial composition. Similarly, rs4145930-G associated with *Bifidobacterium animalis* abundance implicates the gene *HKDC1* in glucose metabolism influencing microbiota.

Conclusion

This thesis enhances genomic research in the Japanese population by:

1. Improving imputation quality: Developing a population-specific reference panel (1KG+7K) significantly improved imputation accuracy, leading to the discovery of novel loci associated with serum uric acid and total cholesterol levels.

2. Exploring host genetics-microbiome interactions: Identifying genetic associations with gut microbiome composition and highlighting the impact of methodological differences on reproducibility.