Developing and Assessing a Genotype Imputation Pipeline for the Taiwanese Population Utilizing Multiple Reference Panels

以多元參考模板建立與評估適用臺灣人群的基因型填補流程

Student: Che-Wei Tsai / Supervisor: Yen-Chen Anne Feng

Institute: Department of Public Health, NTU

Background

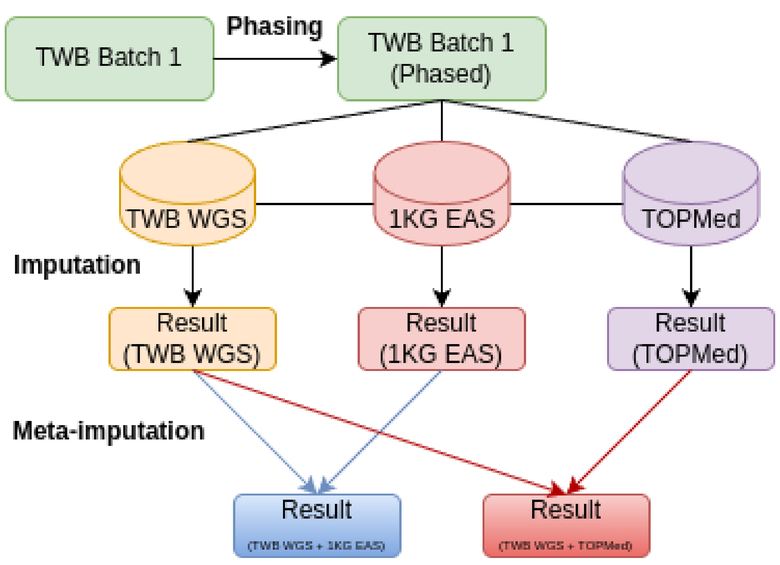
In the realm of genomic research, advancements in sequencing technology have been remarkable. However, when considering large-scale databases like biobanks, conducting whole-genome sequencing (WGS) for all samples still entails significant financial and time investments. Genotype imputation is a process of predicting and imputing genotypes that are not directly assayed in a sample of individuals. This study aims to evaluate the efficacy of genotype imputation on Taiwan Biobank (TWB) individuals using imputation and meta-imputation[2] methods with TWB WGS, 1000 Genomes Project East Asian panel, and TOPMed which consists of samples from diverse ancestry (European ~40%, Hispanic/Latino ~19%, Asian ~8%, and Others ~4%). We also seek to establish a genotype imputation pipeline for the Taiwanese population, facilitating future genetic data analysis for TWB.

Methods

1. Data

- Target sample: TWB Batch 1 (N = 27,274)
- Reference panel:
  - TWB WGS (N = 1,492)
  - 1KG EAS (N = 515)
  - TOPMed (N = 133,597)

2. Flow



3. Tools

	Software	Output
Phasing	Eagle v2.4 [3]	-
Imputation	Minimac4 [4]	R^2, Empirical R^2
Meta-imputation	MetaMinimac2 [5]	Empirical R^2

NOTES:

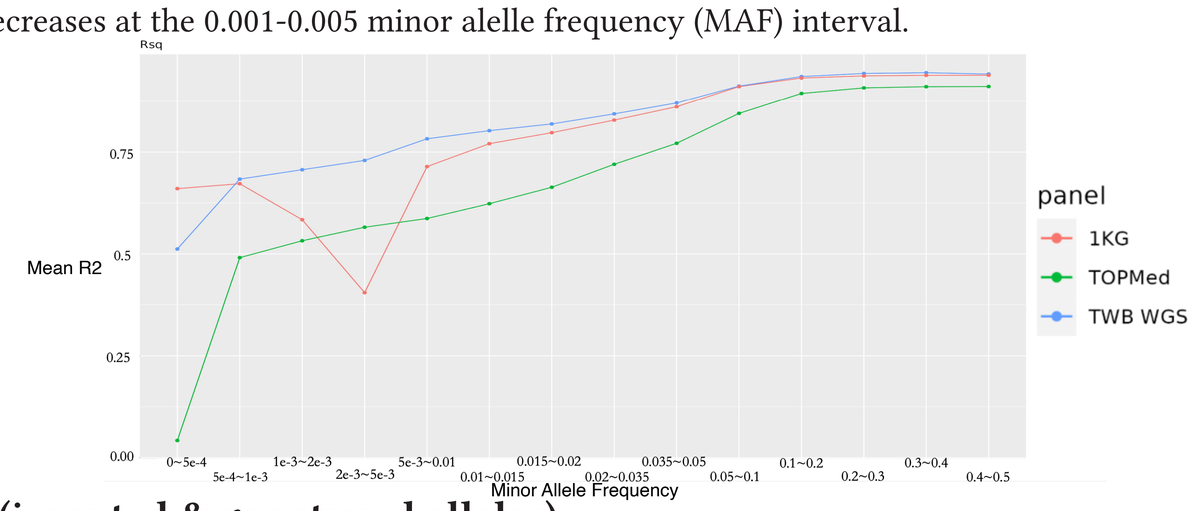
- $R^2 = \frac{\frac{1}{2n} \times \sum_{i=1}^{2n} (D_i - \hat{p})^2}{\hat{p}(1-\hat{p})}$
  - D_i: imputed alternate allele probability
  - $\hat{p}$: alternate allele frequency
  - n: number of sample
- *Empirical* R^2: Pearson squared correlation
- The result of **TWB WGS + TOPMed** hasn't been completed yet.

Results

Since not every allele is genotyped or imputed simultaneously, Minimac4 outputs two different R^2 values to evaluate imputation accuracy for all alleles. R^2, calculated exclusively from imputed dosages, applies to all imputed alleles. Empirical R^2, based on comparisons with actually genotyped markers, applies to both genotyped and imputed alleles.

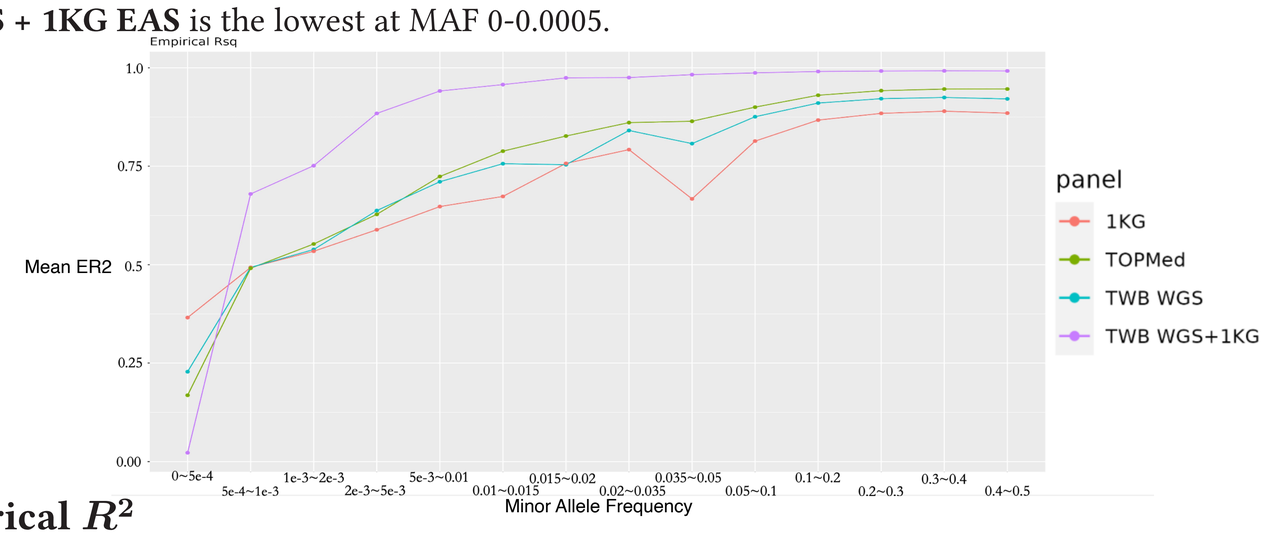
R^2 (imputed alleles)

- **TWB WGS** > **1KG EAS** > **TOPMed**
- **1KG EAS** decreases at the 0.001-0.005 minor alelle frequency (MAF) interval.



Empirical R^2 (imputed & genotyped alleles)

- **TWB WGS + 1KG EAS** > **TOPMed** > **TWB WGS** > **1KG EAS**
- **TWB WGS + 1KG EAS** is the lowest at MAF 0-0.0005.



R^2 v.s. Empirical R^2

- Results are not consistent between R^2 & empirical R^2. (R^2 tends to be larger than empirical R^2)

R2

ER2

Discussion

The properties of reference panel, sample size, and minor allele frequency will all influence imputation performance[1]. According to R^2 results above, we can discover the homogeneity of ancestry in reference panel may also play a key role. The presence of diverse samples in a reference panel (TOPMed) may dilute allele frequency estimates, which are used as prior information in imputation to calculate D_i. Additionally, the potential reason of the decreasing R^2 observed for 1KG EAS at MAF 0.001-0.005 is that only ~0.5% of variants are present within that interval on the 1KG EAS panel.

For empirical R^2, results indicate that meta-imputation, incorporating TWB WGS alongside 1KG EAS, significantly enhances imputation accuracy. However, this improvement is not uniform across all variant categories. Only ~1% and ~99% of variants from TWB WGS and 1KG EAS, respectively, were identified as commonly present on both panels at MAF 0-0.0005 and available for meta-imputation This is because meta-imputation combines imputation results from different panels using weights, and common variants provide a basis for estimating them. This type of issue would pose a significant limitation for meta-imputation.

Finally, compared to empirical R^2, R^2 shows notable inflation. Empirical R^2, relying on true genotype to estimate, provides more direct evidence, thus serving as a more suitable indicator for assessing imputation quality.

Conclusion

1) The TWB WGS + 1KG EAS via meta-imputation is currently the optimal choice for genotype imputation in the Taiwanese population.
2) Increasing sample sizes in each reference panel or merging them with samples from other ancestries may help with enhancing TWB genotype imputation accuracy in the future, especially for rare variants. Moreover, our results suggest that sample sizes between the two panels should be similar to ensure there're sufficient common rare variants in two different panels.

References:

1. *Marchini et al. 2010 Nat Rev Genet*
2. *Yu et al. 2022 Am J Hum Genet*
3. *https://github.com/poruloh/Eagle*
4. *https://github.com/statgen/Minimac4*
5. *https://github.com/yukt/MetaMinimac2*