1.3 Task III 1. In the original article, the authors discuss how they used an imputation technique on they're genetic dataset. Briefly explain what is is an imputation technique and why the authors decided to use it in this work.

Imputation in genetics or genetic imputation is a process of predicting genotypes that are not directly assessed in a sample of individuals. It is a statistical inference of unobserved genotypes achieved by using known haplotypes (i.e. a group of alleles that are inherited together from a single parent) in a population. That is, a reference panel of haplotypes at a dense set of single nucleotide polymorphisms (SNP) (as well as indels and structural variants) is used to impute the unknown genotypes into a study sample of individuals who already have a known subset of SNPs assayed. Broadly speaking, the process of imputation can be divided into two main steps, also implemented in the study of the featured article, namely (i) pre-phasing and (ii) imputation [1]. In pre-phasing, a statistical method is applied to the genotype data to infer the underlying haplotypes of each individual (i.e. the samples to be imputed are pre-phased). Then, in the second step, the inferred haplotypes are combined with a reference panel of haplotypes, based on a different statistical method, to impute the unobserved genotypes in each sample.

A number of factors may influence the accuracy of genetic imputation, and apparently the quality of the reference panel of haplotypes has an impact on it. Generally, accuracy increases as the number of haplotypes in the reference panel grows and also if the ancestry of the reference panel haplotypes is a good match to the ancestry of the sample haplotypes. The UK Biobank dataset, from which the samples of the current analysis study was taken, consists of samples with a wide variety of ancestries, but with the majority of them having British (/European) ancestry. To address and match this unique characteristic of the dataset, a merged haplotype reference panel (rather than a single one) was employed in the imputation process here [2]. As a relevant note, a validation experiment was carried out to evaluate the imputation performance on this dataset, with overall satisfactory performance reported [1].

The main advantage for the authors to apply the imputation technique is that it boosts the number of genotypic materials (SNPs) that can be tested for association (or other analysis). This potentially increases the power of the study, improves the ability to fine-map the causal variants, and facilitates future meta-analysis.

Moreover, imputation (including phasing) can be a computationally intensive procedure. The authors decided to carry out the large-scale imputation centrally also means that other different research groups who are planning to use the dataset do not need to carry this out independently, thereby simplifying some processes in future secondary data analysis.

[1] UK Biobank. Genotype imputation and genetic association studies of UK Biobank: interim data release. http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/imputation_documentation_May2015.pdf. Published May 2015. Accessed May 17, 2019.

[2] Bycroft, C., et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562, 7726 (2018), 203-209.