STAT840 Report

Genotype imputation accuracy evaluation with fastPHASE in TNF gene

Minzhi Huang

December 27th

Abstract

This project investigated the relationship between fastPHASE imputation accuracy and missing genotype rate in a small reference panel. I phased the haplotypes, imputed the genotypes, and evaluated the imputation accuracy of fastPHASE with Tumor necrosis factor gene (reference panel, see https://pga.gs.washington.edu/data/tnf/) from SeattleSNPs. The analysis concluded the fastPHASE imputation accuracy rate does decrease with the increasing missing genotype rate in a small reference panel. Further studies need for the generalization and verification of this conclusion.

Keywords: GWAS, Genotype Imputation, fastPHASE, Tumor Necrosis Factor

1 Introduction

Genome-wide association study is a crucial research method for finding genetic factors related to complex diseases such as cancer, diabetes, heart disease, and mental illness[1][2][3]. Genotype imputation is a widely used and important tool in Genome-wide association studies. A proper evaluation of current popular imputation software's accuracy can boost the power of genome-wide association studies and help researchers to study complex diseases. In this report, I evaluated the imputation accuracy of fastPHASE with various haplotype missing rates (P=10% to P=90%, repeat=5). The result of imputation accuracy is provided in Table 1.

2 Methodology

2.1 Dataset: SeattleSNPs and TNF

The SeattleSNPs database contains common SNPs in more than 100 genes with the study of 24 European Americans and 24 African Americans [4]. Human Leukocyte Antigen is the gene that encodes the human major histocompatibility complex (MHC) and is located on the short arm of chromosome 6 (6p21.31). The TNF loci are approximately 220 kb centromeric to the HLA-B locus in the class III region of MHC [5]. The TNF gene plays an important role in inflammatory conditions. In this

report, I considered sequenced TNF gene data on SeattleSNPs, containing 47 PGA Sample IDs and 24 SNPs sites. The downloaded prettybased dataset is converted into FastPHASE read-in ready format in R with 1 stand for more frequency allele and 0 stand for less frequency allele. All missing SNPs are replaced by question marks ("N" in original prettybase dataset). Each sample with 24 SNPs sites and one genotype is documented in two rows. In the FastPHASE read-in ready data, the first row and the second row are two haplotypes for sample D001 with 24 SNPs sites.

2.2 Haplotype estimation and genotype imputation

The samples that have genotype missing rates larger than 15% (calculated in R) are removed, resulting in removing PGA Sample IDs D001, D005, D009, D014, E001, E016. The remaining genotypes were estimated using fastPHASE 1.4.8 [6]. For each genotype, I masked 10% to 90% of the SNPs at random, which resulted in 98-886 missing genotypes. The error rate was calculated as the proportion of masked genotypes that were not estimated correctly.

3 Results

The results for fastPHASE genotype imputation error rates in Table 1 were all obtained by averaging imputation errors (repeation=5) estimates from T = 20 starts of the EM algorithm.

Command: ./fastPHASE -n -T20 -output -input file name

Table 1: Error Rates for Estimation of Missing Genotypes

FastPhase	**Error Rate for Genotype Imputation (SeattleSNPs TNF)**								
Mask Rates	***10% Masked***	***20% Masked***	***30% Masked***	***40% Masked***	***50% Masked***	60% Masked	***70% Masked***	***80% Masked***	***90% Masked***
Imputation 1	**0.3%**	**1.22%**	**2.03%**	**2.44%**	**7.72%**	**9.35%**	**15.55%**	**10.26%**	**13.82%**
Imputation 2	**0.3%**	**0.5%**	**1.73%**	**2.95%**	**5.08%**	**9.76%**	**9.15%**	**10.77%**	**17.78%**
Imputation 3	**0.3%**	**1.02%**	**1.83%**	**2.85%**	**7.01%**	**7.42%**	**14.3%**	**7.31%**	**20.22%**
Imputation 4	**0.3%**	**1.73%**	**1.93%**	**3.76%**	**5.39%**	**10.78%**	**5.89%**	**7.52%**	**13.52%**
Imputation 5	**0.5%**	**1.22%**	**2.13%**	**2.95%**	**7.11%**	**5.69%**	**8.74%**	**12%**	**10.57%**
Average Error Rate	**0.34%**	**1.4%**	**1.93%**	**2.99%**	**6.46%**	**8.6%**	**10.41%**	**9.57%**	**15.18%**

The error rates range from 0.3% to 20.22% with the missing genotype rates from 10% to 90%. The largest imputation error differences happen at P=70% and P=90% with differences of 9.66% and 9.65%.

4 Discussion

Notwithstanding its limitation, the analysis does show the imputation accuracy rate does decrease with the increasing missing genotype rate. There are many comprehensive reviews of current popular imputation tools with various genetic datasets in the statistical genetics field. The concluding result in this report need to be verified

with a larger reference panel, genotype data quality control, and consideration of other factors which affect imputation accuracy in the future studies[7].

5 Future Work

5.1 Data Quality Control

Data quality assessment and control could be implemented by PLINK before genotypes imputation [8]. Remove the markers had a greater than X% overall missing rate, and had a MAF of less than X% [9].

5.2 Reference Panels

A larger Reference Panels, different Human Genome Projects (1000 Genomes Project, HapMap, etc.), or different gene SNPs (HLA-A, HLA-C, HLA-DP, HLA-DM, etc.) could be used in imputation process.

5.3 Imputation Accuracy Measurement

The imputation repetition could be increased (n=50) and taking the average of imputation accuracy as the accessed imputation accuracy. The average squared correlation between masked array genotypes and imputed alleles could be added to the imputation accuracy assessment[10].

5.4 Additional

1. Imputation accuracy assessment could be done with multiple current popular imputation tools (MaCH, BEAGLE, IMPUTE2, SHAPEIT, etc,).
2. The model algorithm for haplotypes inferring and genotypes inferring could be discussed in the Methodology section.

6 Appendix

6.1 Sites

000214, 000282, 000346, 000352, 000657, 001009, 001078, 001278, 001440, 001476, 001893, 002642, 002984, 003018, 003746, 004013, 004040, 004101, 004205, 004231, 004366, 004671, 004711, 004765.

References

[1] C. C. Chung, W. C. Magalhaes, J. Gonzalez-Bosquet, and S. J. Chanock, "Genome-wide association studies in cancer—current and future directions," Carcinogenesis, vol. 31, no. 1, pp. 111–120, 2010.

[2] C. J. Willer, E. M. Schmidt, S. Sengupta, G. M. Peloso, S. Gustafsson, S. Kanoni, A. Ganna, J. Chen, M. L. Buchkovich, S. Mora et al., "Discovery and refinement of loci associated with lipid levels," Nature genetics, vol. 45, no. 11, p. 1274, 2013.

[3] S. P. Hagenaars, S. E. Harris, G. Davies, W. D. Hill, D. C. Liewald, S. J. Ritchie, R. E. Marioni, C. Fawns-Ritchie, B. Cullen, R. Malik et al., "Shared genetic aetiology between cognitive functions and physical and mental health in uk biobank (n= 112 151) and 24 gwas consortia," Molecular psychiatry, vol. 21, no. 11, pp. 1624–1632, 2016.

[4] E. Jorgenson and J. S. Witte, "Coverage and power in genomewide association studies," The American Journal of Human Genetics, vol. 78, no. 5, pp. 884–888, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0002929707638221

[5] P. Posch, I. Cruz, D. Bradshaw, and B. Medhekar, "Novel polymorphisms and the definition of promoter 'alleles' of the tumor necrosis factor and lymphotoxin α loci: inclusion in hla haplotypes," Genes & Immunity, vol. 4, no. 8, pp. 547–558, 2003.

[6] P. Scheet and M. Stephens, "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase," The American Journal of Human Genetics, vol. 78, no. 4, pp. 629–644, 2006.

[7] S. Das, G. R. Abecasis, and B. L. Browning, "Genotype imputation from large reference panels," Annual review of genomics and human genetics, vol. 19, pp. 73–96, 2018.

[8] C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan, "Data quality control in genetic case-control association studies," Nature protocols, vol. 5, no. 9, pp. 1564–1573, 2010.

[9] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell et al., "The uk biobank resource with deep phenotyping and genomic data," Nature, vol. 562, no. 7726, pp. 203–209, 2018.

[10] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis, "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing," Nature genetics, vol. 44, no. 8, pp. 955–959, 2012.