Imputation from low-pass whole genome sequencing data with GLIMPSE2 versus TOPMed

Zinhle Cindi[1], Yuki Bradford[1], Phumla Sinxadi[2,3], David W. Haas[4,5], Marylyn D. Ritchie[1,6]

[1]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Division of Clinical Pharmacology, Department of Medicine, University of Cape Town, Cape Town, South Africa
[3]SAMRC/UCT Platform for Pharmacogenomics Research and Translation, South African Medical Research Council, Cape Town, South Africa
[4]Vanderbilt University Medical Center, Nashville, TN 37232, USA
[5]Meharry Medical College, Nashville, TN 37208, USA
[6]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

Background: DNA sequencing technologies are emerging as alternatives to genotyping arrays. Sequencing enables comprehensive probing of the entire genome with a specified depth of coverage. Whole-genome sequencing (WGS) performed with a read depth of 30x to 50x is presently cost-prohibitive for most large-scale studies. Low-pass whole-genome sequencing (lp-WGS) has emerged as a cost-effective alternative to genotyping arrays and high depth WGS. lp-WGS typically involves a read depth of 0.1x to 1x, with subsequent imputation. Imputation from genotyping arrays has been widely used in large-scale genome-wide association studies (GWAS), using either 1000 Genomes or TOPMed reference panels. GLIMPSE2 was specifically designed to impute from lp-WGS data. It is unclear whether standard GWAS imputation approaches could be applied to lp-WGS data, or whether other pipelines such as GLIMPSE2 are required. The present methods analyses compared genomic coverage and data quality imputed from lp-WGS using GLIMPSE2 versus TOPMed.

Methods: Using a test set of lp-WGS data generated from five DNA samples, we performed imputation using both GLIMPSE2 and TOPMed. Unfiltered data were submitted to imputation pipelines. Concordance between output variants was assessed. We also checked for concordance between pre-imputation variant calls with GLIMPSE2 versus TOPMed.

Results: There was 96.0% concordance among variants shared by datasets imputed by both GLIMPSE2 and TOPMed. GLIMPSE2 generated 63,824,182 variant calls while TOPMed generated 8,439,092 variant calls. Concordance between pre-imputation and imputed variants was 88.3% with GLIMPSE2 (n=11,950,240 non-missing calls) and 85.1% with TOPMed (n=6,379,129 non-missing calls).

Conclusions: Imputation of lp-WGS data with GLIMPSE2 yielded 7.6-times more variants than with TOPMed, indicating that TOPMed is not suitable for lp-WGS data. There was good agreement between variants generated by the two methods. It will be important to compare lp-WGS variant calls imputed with GLIMPSE2 versus variant calls generated using other standard methods.